2012-06

Application and Survey of Business Intelligence (BI) Tools within the Context of Military Decision Making

Tounsi, Mohamed Ilyes

Monterey, California. Naval Postgraduate School

NAVAL POSTGRADUATE SCHOOL

MONTEREY, CALIFORNIA

THESIS

APPLICATION AND SURVEY OF BUSINESS INTELLIGENCE (BI) TOOLS WITHIN THE CONTEXT OF MILITARY DECISION MAKING

by

Mohamed Ilyes Tounsi

June 2012

Thesis Advisor:	Magdi Kamel
Second Reader:	Walter Kendall

Approved for public release; distribution is unlimited

THIS PAGE INTENTIONALLY LEFT BLANK

REPORT DOCUMENTATION PAGE			Form Approved OMB No. 0704-0188
Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instruction, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188) Washington DC 20503.			
1. AGENCY USE ONLY ***(Leave blank)***	**2. REPORT DATE** June 2012	**3. REPORT TYPE AND DATES COVERED** Master's Thesis	
4. TITLE AND SUBTITLE Application and Survey of Business Intelligence (BI) Tools within the Context of Military Decision Making		**5. FUNDING NUMBERS**	
6. AUTHOR(S) Mohamed Ilyes Tounsi			
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Naval Postgraduate School Monterey, CA 93943-5000		**8. PERFORMING ORGANIZATION REPORT NUMBER**	
9. SPONSORING /MONITORING AGENCY NAME(S) AND ADDRESS(ES) N/A		**10. SPONSORING/MONITORING AGENCY REPORT NUMBER**	
11. SUPPLEMENTARY NOTES The views expressed in this thesis are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government. IRB Protocol number _____N/A_____.			
12a. DISTRIBUTION / AVAILABILITY STATEMENT Approved for public release; distribution is unlimited		**12b. DISTRIBUTION CODE**	

13. ABSTRACT (maximum 200 words)

Business intelligence (BI) is a general category of applications and technologies for collecting, storing, analyzing, and providing access to data to help users make better and faster decisions. BI applications include the activities of decision support systems, query and reporting, online analytical processing (OLAP), statistical analysis, forecasting, and data mining. The purpose of this research is to explore and survey several tools that fall under the BI umbrella and investigate their applicability within the context of military decision making. This survey will help military decision makers select the right BI tool for the right decision problem using the right technology. This would result in reduced IT costs by eliminating redundancy and consolidating computing resources, accelerated decision making, and improved accuracy, consistency, and relevance of decisions by providing a single version of truth.

14. SUBJECT TERMS Business Intelligence, OLAP, OLTP, OBIEE, Rapid-I, PolyAnalyst, Oracle, decision making, data warehouse			**15. NUMBER OF PAGES** 103
			16. PRICE CODE
17. SECURITY CLASSIFICATION OF REPORT Unclassified	**18. SECURITY CLASSIFICATION OF THIS PAGE** Unclassified	**19. SECURITY CLASSIFICATION OF ABSTRACT** Unclassified	**20. LIMITATION OF ABSTRACT** UU

NSN 7540-01-280-5500

Standard Form 298 (Rev. 2-89)
Prescribed by ANSI Std. 239-18

THIS PAGE INTENTIONALLY LEFT BLANK

Approved for public release; distribution is unlimited

APPLICATION AND SURVEY OF BUSINESS INTELLIGENCE (BI) TOOLS WITHIN THE CONTEXT OF MILITARY DECISION MAKING

Mohamed Ilyes Tounsi
Captain, Tunisian Air Force
Communication Engineer, Tunisian Air Force Academy, 2001

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN SYSTEMS TECHNOLOGY

from the

NAVAL POSTGRADUATE SCHOOL
June 2012

Author: Mohamed Ilyes Tounsi

Approved by: Associate Professor Magdi Kamel
Thesis Advisor

Lecturer Walter Kendall
Second Reader

Professor Dan Boger
Chair, Department of Information Sciences

THIS PAGE INTENTIONALLY LEFT BLANK

ABSTRACT

Business intelligence (BI) is a general category of applications and technologies for collecting, storing, analyzing, and providing access to data to help users make better and faster decisions. BI applications include the activities of decision support systems, query and reporting, online analytical processing (OLAP), statistical analysis, forecasting, and data mining. The purpose of this research is to explore and survey several tools that fall under the BI umbrella and investigate their applicability within the context of military decision making. This survey will help military decision makers select the right BI tool for the right decision problem using the right technology. This would result in reduced IT costs by eliminating redundancy and consolidating computing resources, accelerated decision making, and improved accuracy, consistency, and relevance of decisions by providing a single version of truth.

THIS PAGE INTENTIONALLY LEFT BLANK

TABLE OF CONTENTS

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF FIGURES

LIST OF TABLES

THIS PAGE INTENTIONALLY LEFT BLANK

ACKNOWLEDGMENTS

This thesis, a result of my two years of graduate studies at the Naval Postgraduate School, is dedicated to my parents and my parents-in-law who encouraged me at every step of my life; to my brothers and brothers-in-law who had more confidence in me than I had in myself; to my two sons, Ahmed and Mohamed Anass, who supported me with their love; and, most importantly, to my dear wife, Wafa, who unflaggingly supported and tolerated me while I prepared the research project.

I would also like to express my sincerest thanks to all the people who have helped me along the way in preparation of this thesis. I would particularly like to thank my outstanding advisors; Associate Professor Magdi Kamel and Walter Kendall, for their patience, assistance and guidance in helping me prepare this thesis.

Finally, I sincerely thank all NPS staff for providing me with the necessary help. Without all of your contributions this would not have been possible. Thank you.

THIS PAGE INTENTIONALLY LEFT BLANK

I. INTRODUCTION

Business intelligence is a general category of methodologies, technologies, and applications for collecting, storing, analyzing, and providing access to data to help users make better and faster decisions. The word *analytics* is often used to describe business intelligence. BI applications include the activities of decision support systems, query and reporting, online analytical processing (OLAP), statistical analysis, forecasting, and data and text mining. The purpose of this research is to explore and survey BI tools within the context of military decision making.

A. SCOPE OF THE THESIS

The scope of the thesis is in the application and survey of business intelligence (BI) Tools within the context of military decision making. It investigates a variety of state of the art BI tools such as Oracle BI tools, Megaputer PolyAnalyst, and Rapid-I data and text mining suites.

This thesis is unclassified. It does not involve any human subject research.

B. PROBLEM STATEMENT

Military decision makers must respond quickly to changing conditions and be innovative in decisions making. Increasingly, military decision makers are turning to computerized support to help them make better decisions. In this thesis I investigate the use of business intelligence tools to improve decision making in the context of military applications.

C. PURPOSE STATEMENT

The purpose of this thesis is to explore and survey the use of BI and BI tools in military decision making. This research will allow the defense establishment to use powerful tools to accelerate decision making, and improve its accuracy, consistency, and relevance.

D. LITERATURE REVIEW

Literature review consisted of review of books, articles, websites, and other BI resources.

E. RESEARCH QUESTIONS

In the performance of this study, this thesis will address the following questions:

1. What is BI and how can it support decision making in military organizations?
2. What are some of the state of art BI tools (open source and commercial) and their capabilities?
3. How BI applications are implemented and deployed?
4. What are the expectations of BI tool implementation and deployment in support of military decision makers?

F. RESEARCH METHODS

In order to achieve the thesis goals, a review of existing research in implementing and deploying BI tools such as querying, reporting, data mining, and others was performed. Example data were used to explore the capabilities of a number of BI tools in term of delivering greater insight to end users in organizations through the use of dashboards, query, analysis and alerts, data and text mining, as well as other analytics.

G. PROPOSED DATA, OBSERVATION AND ANALYSIS METHODS

BI tools will be analyzed for cost savings and timeliness of decisions in military decision making.

H. POTENTIAL BENEFITS, LIMITATIONS AND RECOMMENDATIONS

The benefits of using appropriate BI tools as identified by this research will help:

- Reduce IT costs by defining the right problem.

- Accelerate decision making by being able to quickly and easily create reports and queries, embed business intelligence, and rapidly model alternative scenarios.
- Improve accuracy, consistency, and relevance of decisions by providing a single version of truth.

I. CHAPTERS OUTLINE

This thesis is divided into six chapters. Chapter II presents and overview of business intelligence concepts, technologies, and architectures. Chapter III describes PolyAnalyst, a commercial data and text mining package from Megaputer Intelligence, Inc. Chapter IV presents an overview of Oracle BI, a set of business intelligence tools from Oracle Corporation. Chapter V depicts the processing and analysis capabilities of RapidMiner, an open source system for knowledge discovery and data mining by Rapid-i. Finally, Chapter VI presents a summary, conclusions, and recommendations for future work.

THIS PAGE INTENTIONALLY LEFT BLANK

II. BUSINESS INTELLIGENCE CONCEPTS

This chapter introduces and defines Business Intelligence concepts and frameworks. It will serve as a reference to BI products presented in next chapters three, four, and five. It begins by defining BI components which are the data warehouse, BI analytics, and the business performance management. Then, it describes the data warehousing architecture and process. Then it illustrates a review on business analytics and the Online Analytic Process (OLAP). Afterwards, it introduces data, text and web Mining techniques and tools. Finally, the chapter wraps up with an overview of performance measurement system.

A. INTRODUCTION TO BUSINESS INTELLIGENCE

1. What Is Business Intelligence?

Business intelligence (BI) is a set of methodologies and technologies that enable business managers to access data, manipulate it, and conduct analysis for decision making. BI has three main components. The first is a data warehouse, which is a repository that collects data from multiple sources and organizes it for decision making. When the amount of data is really huge with the high rate of analysis processing, online business transactions may be slowed down if a separate data warehouse is not used with BI. A data warehouse can be replaced by a data mart for small companies with small amount of data and simple data analytics. The second component is BI tools or analytics and visualization for manipulating, mining, and analyzing the data in the data warehouse. The third and final component is the business performance management (BMP) for monitoring and analyzing the performance of the organization [2]. A user interface enables users to interact with the BI and business performance management tools.

2. Theories and Characteristics of Business Intelligence

Today's operational systems collect data from day-to-day transactions—bank deposits, ATM withdrawals, cash register scans at the store, etc. These Transaction Processing systems are persistently performing updates to Operational Databases. These

systems, called online transaction processing systems (OLTP), handle an organization's routine ongoing business. A data warehouse, on the other hand, is a repository that allows analysis of data for decision making. A data warehouse collects data for online analytic processing (OLAP), organizes it, and enables the user to query it to conduct analysis.

The following are some metaphors and approaches of BI:

- A factory and warehouse. In this view, a data warehouse is viewed as a model of a factory, receiving materials from warehouses and distributing products back to the market place [1].
- The information factory, as shown in Figure 1, is moving toward the web environment. Similar to a factory, the information factory utilizes data sources as inputs, DW and datamarts as storage, analysis and data mining as input processing, and data delivery and BI applications as outputs [2].

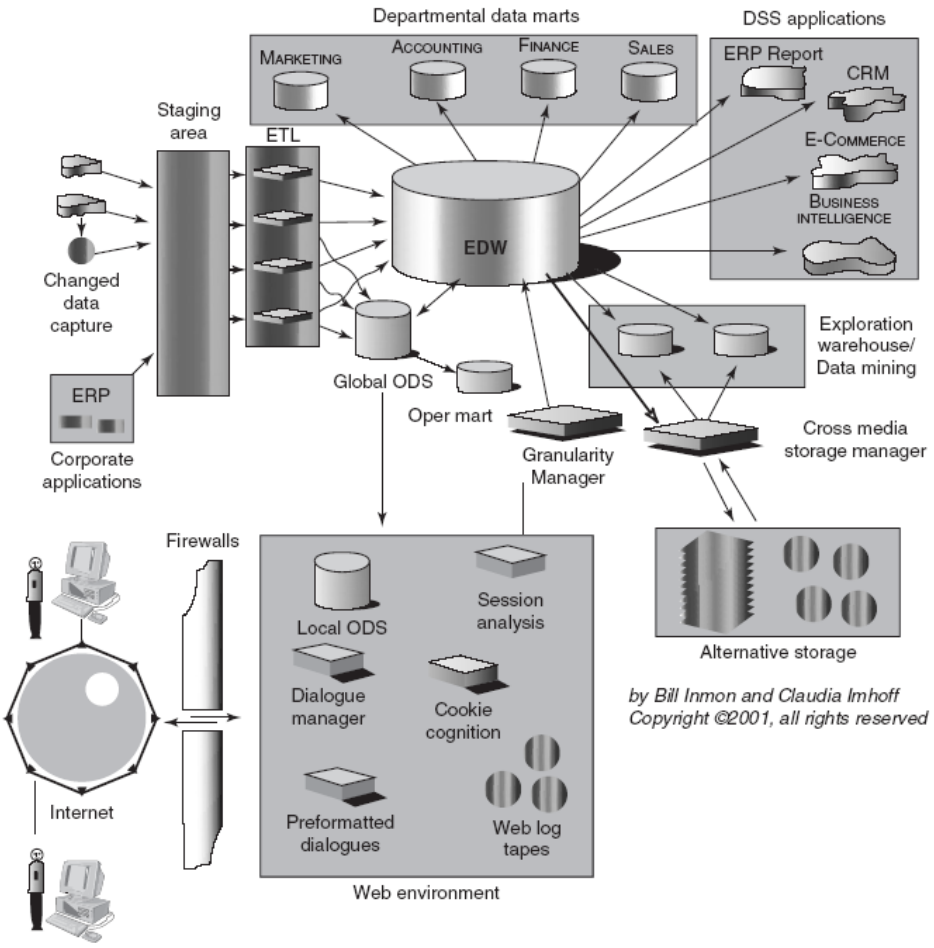


Figure 1 The corporate Information Factory. (From [1])

- Data warehousing and business intelligence. A data warehouse is a set of data used for reporting and analysis helping with decision making. It contains a mixture of selected data that represents a picture of business conditions at a specific moment in time. The main issue in this view is to create relevant data out of OLTP systems in such efficient manner for querying, analysis, and then decision making [2].
- Teradata advanced analytics methodology. As shown in Figure 2, Teradata created a different approach for BI. This methodology provides a complete set of techniques that allows building new models, create new views of data, generate simulation scenarios, and assist not only in understand realities but also predicting result of future states [2].

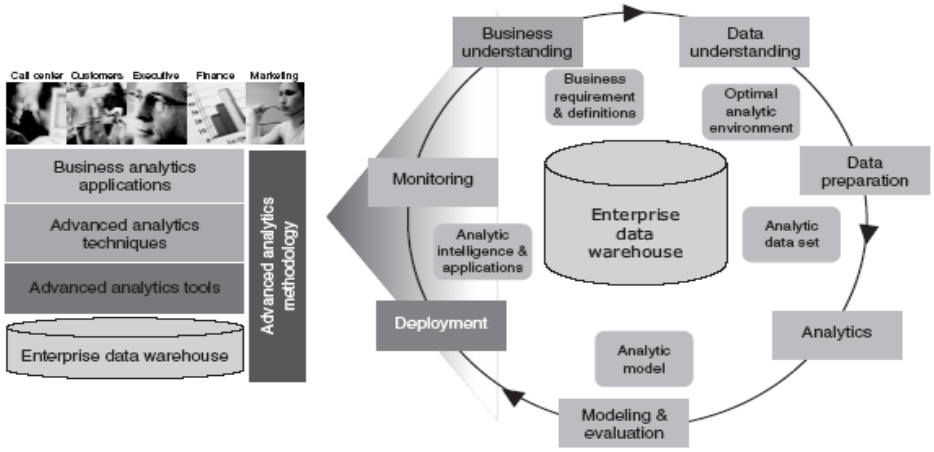


Figure 2 Teradata Advanced Analytics Methodology. (From [2])

- Oracle BI system. Oracle Inc. is recognized for being a specialized vendor in integrating databases and analysis. Figure 3 shows Oracle structural methodology. The methodology illustrates the BI contribution to achieve the enterprise strategic advantage [2].

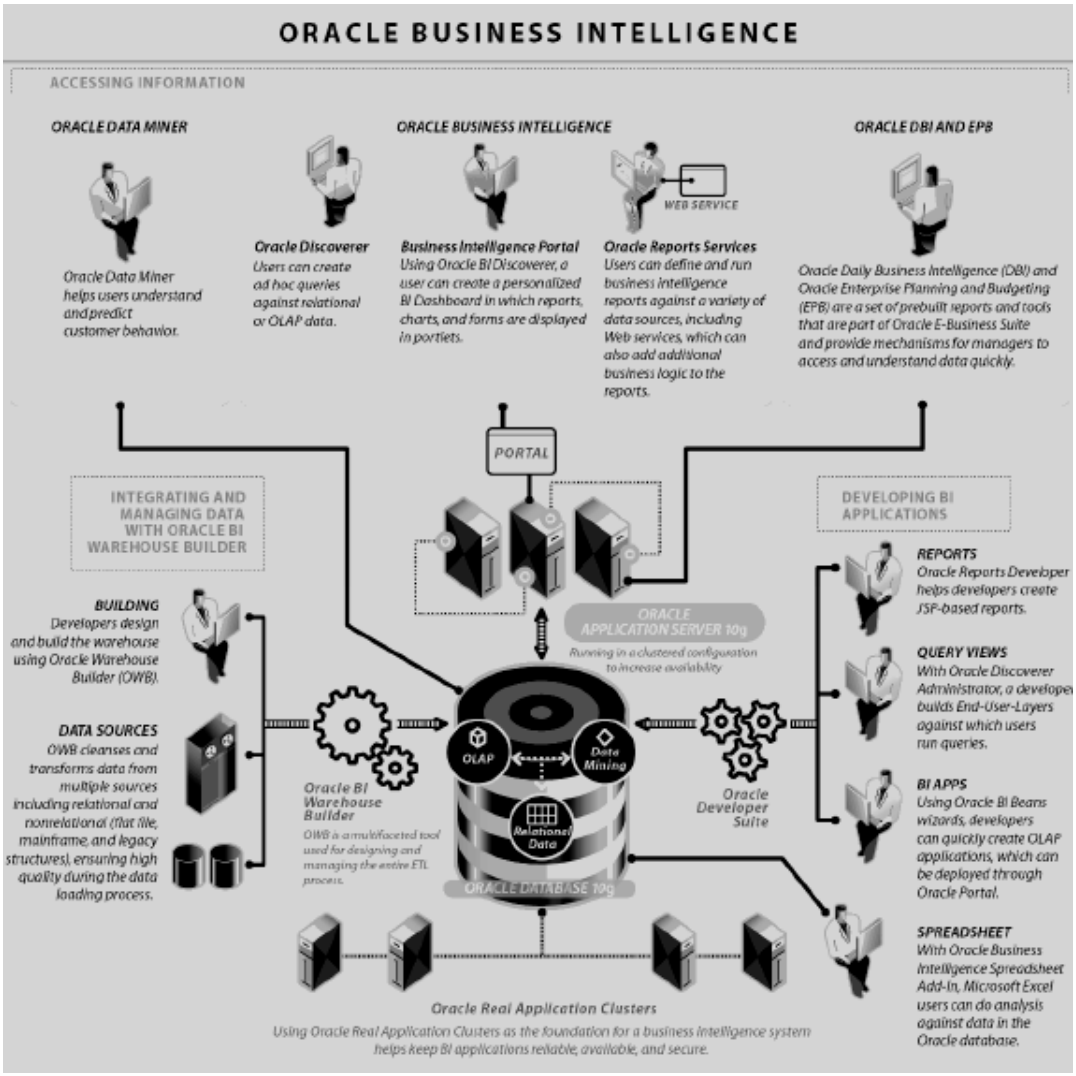


Figure 3 The Oracle BI system. (From [2])

3. Benefits of BI

Today's companies such as Google, Amazon.com, Apple, and even Walmart and Target increasingly rely on BI to achieve a competitive advantage. Companies make use of the huge amount of information available to them and maximize the use of their data assets [2]. In the military domain, a good BI product with accurate data warehouse allows the commander to quickly select the best people for the appropriate mission. Dashboards and other visualization tools are employed by producers, retailers, and other companies. Several analytical tools are used to assist and facilitate decision making whatever the user

level is. A successful BI system achieves considerable benefits for the enterprise. A survey of 510 corporations, performed by Ekerson (2003), indicates BI benefits as follow [2], [5]:

- Time saving (61 percent)
- Single version of truth (59 percent)
- Improved strategies and plans (57 percent)
- Improved tactical decisions (56 percent)
- More efficient processes (55 percent)
- Cost savings (37 percent)

Additional benefits include the ability of BI systems to provide accurate information when needed, including real-time performance analysis support decision making for strategic planning [2].

"The Trac2es system, Transcom Regulating and Command and Control Evacuation System, includes decision support, reporting, and analysis tools used for tracking and coordinating movement of patients for medical care", said Lt. Col. Keith Lostroh, functional program manager of Trac2es [3]. U.S. military are using business intelligence tools to track evacuated injured soldiers from battle in Iraq and Afghanistan. BI tools help military personnel making the best decision based on the patient critical information such as clinical data and severity of the injury along with demographic and geographic data to decide, if it is safe, how quickly the patient need to fly to another care facility. In the meantime, the tools provide to that facility the current state of the in patient in order to prepare the appropriate medical stuff [3].

B. DATA WAREHOUSING

1. Data Warehousing Definitions and Concepts

A data warehouse is a collection of current or historical data generated in essence to assist decision making. If the data warehouse is wisely optimized, well organized, and accurately built then it will serve as an efficient platform for BI. This repository is

designed to facilitate analytical processing activities such as OLAP, data mining, querying, reporting, and any other decision support processes.

Data warehouse data has a number of characteristics:

- subject oriented
- categorized by subjects containing relevant information for decision support
- details like units and measures are integrated in context of naming conflicts and differences time variance of historical data is preserved in order to perceive trends, deviations, long-term relationships for anticipation and comparisons. These data are characterized this way to improve decision making.
- not alterable by users, i.e., data is nonvolatile.

Along with the previous characteristics, a data warehouse is distinguished by being a web based, multidimensional, client-server, real time, and includes metadata, which describe the structure and the semantics of the data.

A data mart is a smaller and more focused version of a Data Warehouse. It can be viewed as a subset of a Data Warehouse that deals with only one particular topic or area of interest. A dependent data mart is extracted directly from a data warehouse. An independent data mart is miniature warehouse created for a strategic business unit (SBU).

An operational data store (ODS) is similar to a customer information file (CIF). Its content is maintained during the business operations. ODS is applied in short-term mission decisions applications.

2. Data Warehousing Process Overview

As shown in Figure 4, organizations perform their own data warehousing process. Data is extracted (copied) from various external sources selected, cleansed, transformed to a specified format, and integrated according to their decision application model. Subsequently, and according to specific organizational areas such as marketing, Risk

management, or Engineering, data marts are loaded from the populated data warehouse. Analysts often utilize middleware to access the data warehouse. They may create their own SQL queries or employ a managed query environment like Business Objects. The front end users may utilize various applications for data mining, OLAP, reporting and visualization tools.

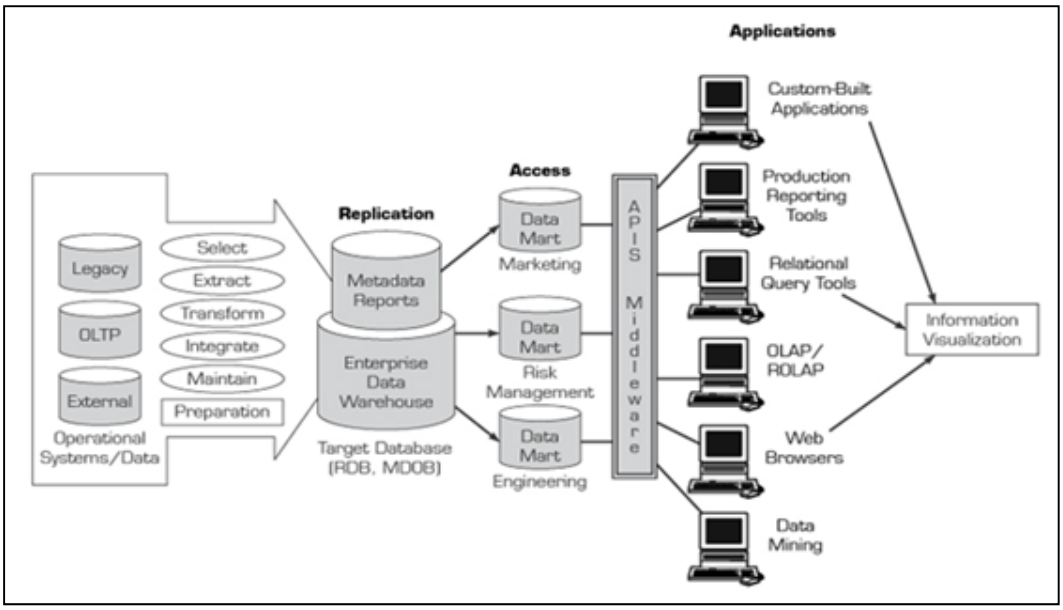


Figure 4 Data warehouse framework and views. (From [2])

3. Data Warehousing Architectures

There are two data warehousing architectures commonly in use: two-tier and three-tier architecture. The data warehousing environment can be broken down as follows:

1. The data warehouse itself that consists of the data and associated software.
2. Data acquisition (back-end) software, which performs the ETL process leading to loading the data into the data warehouse.
3. Client (front-end) software (analytics tools), which allows end users to access and analyze data in the warehouse.

In a three-tier architecture, all the three components of the data warehousing environment coexist. The advantage of this design is the separation of functions, which simplify the creation of data marts. The two-tier architecture represents an economical design because both the data warehouse and the analytics tools run on the same servers.

4. Data Integration and the Extraction, Transformation, and Loading Process

In order for data to fit properly with the data warehouse environment and to enable the process of ETL, data integration must follow three processes: data access capability from any data source, data federation when integrating business views, and change capture which includes identification capture and delivery of all stored changes.

The ETL process consists of data extraction from different data sources, data cleansing, data transformation to the form required by the data warehouse, and data loading to the data warehouse.

5. Data Warehousing Development

The data warehouse development begins with a statement of what an organization wants to accomplish from a data warehousing solution. This statement should align with where the company wants to go, why it wants to go there, and what will it do when it gets there. This approach identifies the strategy of the data warehousing solution.

There are several approaches for structuring the data in a data warehouse. The most commonly used data warehouse structure is the star schema. It utilizes dimensional modeling, which is a retrieval-based approach allowing high volume query access. In this model, several dimensional tables surround the central fact table creating a structure that facilitates querying and decision analysis [2]. The star schema is the final result of the extract, transform, and load (ETL) processes used in building a data warehouse characterized by efficient retrieval of business information.

6. Real-Time Data Warehousing

Real-time data warehousing (RDW) or active data warehousing (ADW) consists of loading and supplying data through the data warehouse as soon as it is available. This process is very helpful in tactical decision making [2]. Moreover, it reduces the discontinuity of data flow and help companies to perform real-time analysis on customer data. An online travel agent is a good example of a system that requires real time data to equally serve its customers and suppliers. Such a system needs to display hotel and airlines pricing information in real time, otherwise customers will turn elsewhere.

7. Data Warehouse Administration and Security Issues

In order to reach the company objective, the data warehouse administrator (DWA) should be technically competent in managing sophisticated software, hardware, and state-of-the-art networks. In addition, he should have a good business understanding as well as decision making processes. Most importantly the DWA needs to have excellent interpersonal and communications skills.

Effective security policies in a data warehouse should focus on performing effective corporate and security policies and procedures, implementing logical security procedures and techniques to restrict access, limit physical access to the data center environment, and establish an effective internal control review process with an emphasis on security and privacy [12].

C. BUSINESS ANALYTICS AND DATA VISUALIZATION

1. Overview

Analytics is the science of analysis. Business analytics (BA) is a set of techniques and tools to gather, store, analyze and provide enterprise users access to data in order to perform faster and better decisions [2].

As shown in Figure 5, BA can be divided into three categories: Information and knowledge discovery, Decision support and intelligent systems, and Visualization.

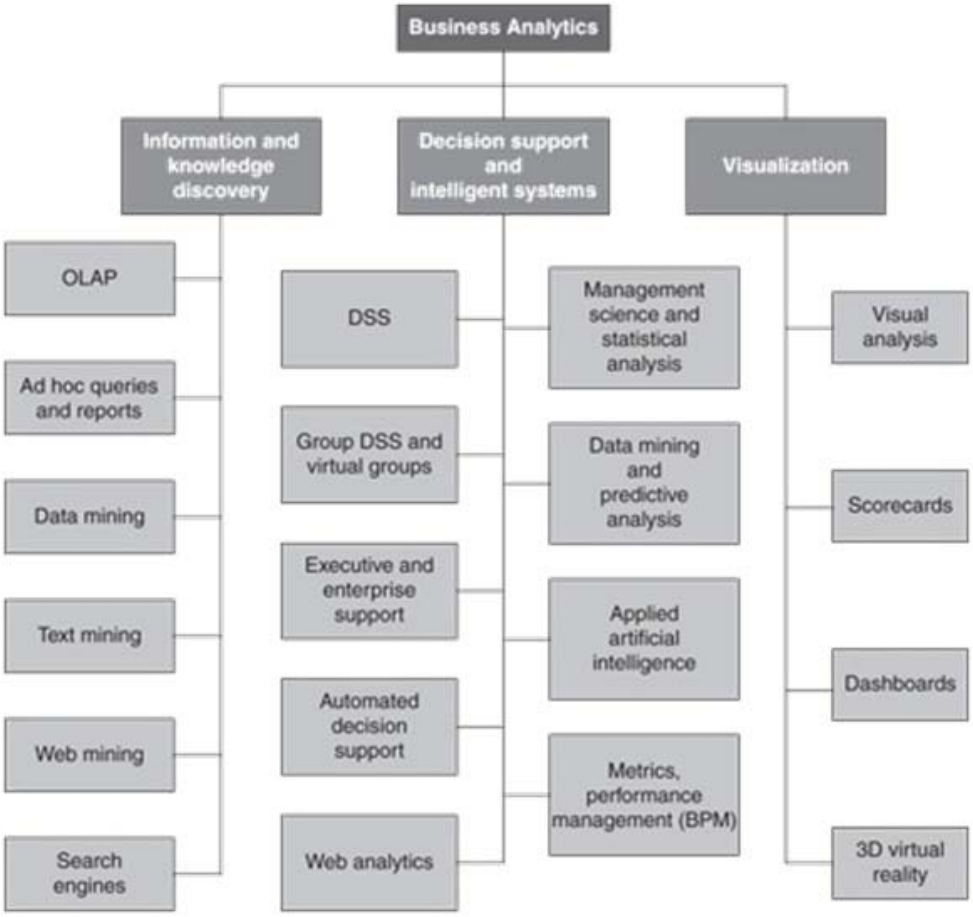


Figure 5 Categories of business analytics. (From [2])

MicroStrategy is a leading BI company that provides integrated reporting, analysis, and monitoring software to help leading organizations make better business decisions every day - on iPad, iPhone, BlackBerry, and more [4]. MicroStrategy classifies its BA product into five styles of BI, as shown in Figure 6. First, enterprise reporting is used to produce static reports that are distributed to numerous end users. These reports are completed using pixel-perfect formats for operational reporting and dashboards. Second, cube analysis performs multidimensional OLAP with slice and dice analytical capabilities. Ad hoc querying and analysis is another BA style that provides relational OLAP ability to query and slice and dice the database as well as drill down capabilities inside the transactional information cube. Statistical analysis and data mining tools are utilized to illustrate cause and effect, correlation, and perform predictive analysis. The

last BA style is report delivery and alerting. This is a proactive application that has the ability to notify a huge population based on subscriptions, schedules, or threshold events.

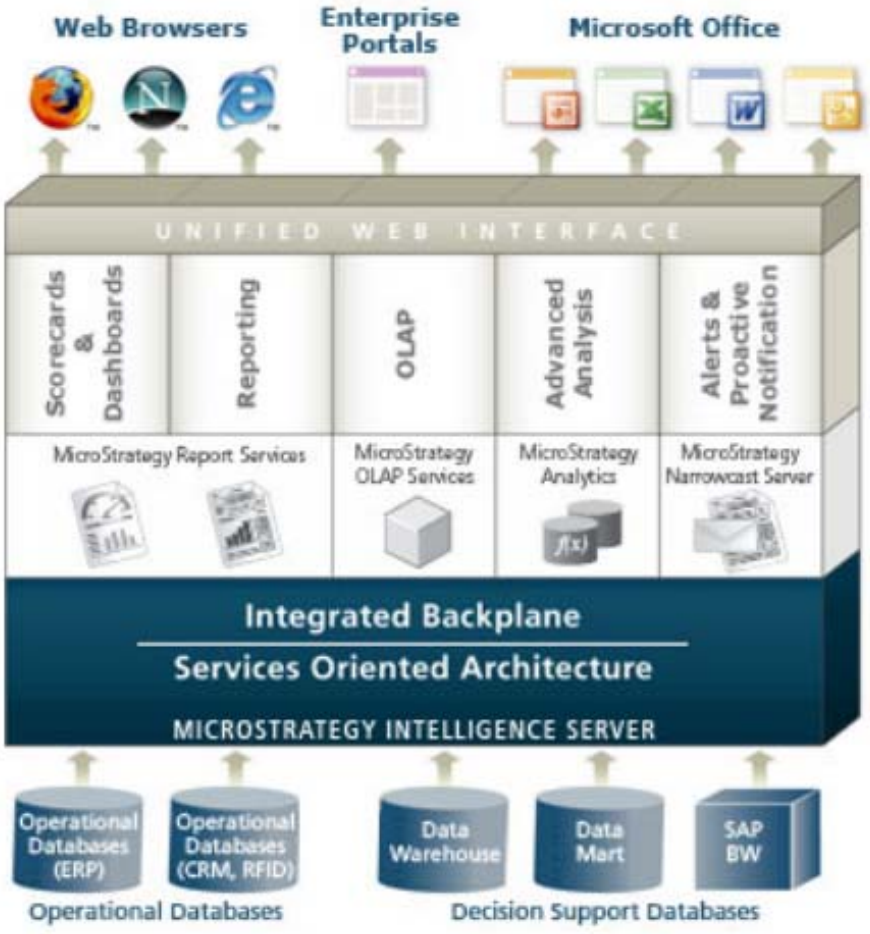


Figure 6 MicroStrategy's platform architecture. (From [8])

2. Online Analytic Process (OLAP)

Online Analytic Process (OLAP) systems are designed for ad hoc analysis and complex queries for data in a data warehouse or data mart providing multidimensional view of the data. While OLTP focuses on large quantity of simple recurring transaction processing, OLAP entails examining thousands and millions of data in complex relationships. OLAP and OLTP depend on each other though since OLAP uses the data generated by OLTP, and OLTP automates the processes that are generated by the decisions supported by OLAP.

OLAP is classified into four types. Multidimensional OLAP or MLAP which includes provisions and apply advanced indexing and hashing when performing queries [13], Relational OLAP or ROLAP in which multidimensional data is stored in a relational database, Database OLAP and web based OLAP, and an economical and simple Desktop OLAP which is defined as the returned result from database executed query as a cube [14].

3. Reports and Queries

A main product of OLAP systems are Queries and Reports. OLAP reports are required to be uniform, flexible, and adjustable. The two major types of reports are routine reports which are periodically and automatically created and distributed, and ad hoc or on-demand reports. Sophisticated systems employ SQL and query by example tools. Some of these systems use intelligent query tools to assist end user in asking the right question.

4. Multidimensionality

The main operational structure of OLAP is a multidimensional one (actual or virtual) that allows for analysis of data. As indicated earlier, a multidimensional structure is characterized by two factors: Dimensions like products, business units, or distribution channels, and Measures like money, sales volume, or forecasted profit

A multidimensional database is a database that supports multidimensional analysis. In a data cube, data is stored according to some measure of interest. A data cube can have two, three, or higher dimensions. Dimensions are attributes in the database, whereas cell contents are the values of interest.

5. Advanced Business Analytics

Today’s organizations are using sophisticated Business Analytics tools including numerous mathematical, financial, and statistical models to uncover problems, discover prospects, boost productivity, and increase outputs.

These advanced techniques include forecasting and estimating models that could lead to improved decision making. Data mining is a special approach that uncovers hidden patterns and relationships in large databases, using machine learning and complex statistical models. Unlike OLAP, data mining is able to answer questions that the users did not think of. These answers could be used to develop predictive models. In Predictive Analysis, models use historical relationships and patterns to predict outcome of new input data.

6. Data Visualization

Data visualization is the use of visual representation to explore, analyze, and interpret large amounts of data. Data visualization is related to all BI technologies. It incorporates graphs, charts, digital images, Geographic Information Systems (GIS), Graphical User Interface (GUI), virtual reality, dimensional presentation, video, and animation. Data visualization facilitates the discovery of relationships and trends, particularly in large amount of data. For example, three-dimensional visualization allows users to easily visualize multiple dimensions of data in a single view.

End user visualization mostly consists of charts and graphs, produced by tools such as Microsoft Excel, in addition to mathematical, statistical, reporting, and querying tools. Top-level executives utilize dashboards and scorecards that incorporate charts, graphs, and tables in a single view.

7. Geographic Information Systems (GIS)

A Geographic Information Systems (GIS) is a system for storing, modeling, integrating, manipulating, and displaying geo-referenced data. By integrating geographically referenced maps in spatial databases, end users can perform many useful tasks that lead to improved decision making. GIS applications are used to improve decision making in the public and private sectors including, dispatch of emergency vehicles, transit management, logistics, facility site selection, and other applications areas. Many leading companies integrated GIS into their BI systems. In fact, Pepsi Cola uses GIS to locate new Taco Bell and Pizza Hut restaurants using traffic patterns [2]. Health insurance is using GIS to select affiliated physicians within a given radius of a

business [2]. Other automobile manufacturers join GIS and GPS to route drivers to destinations efficiently [2].

8. Real-Time Business Intelligence, Automated Decision Support (ADS), and Competitive Intelligence

Traditionally, data warehousing and BI are used to support decision making based on historical data. Today's executives need to perform BI in real or near-real time to respond quickly to a continually changing environment. By implementing automated business process that can be integrated in real time with data in the data warehouse would lead to a timely response to queries, OLAP, and data mining.

The use of real-time data is essential in generating Automated Decision Support (ADS) systems. For instance, in order to approve a loan and grant a credit line to a customer, real time and high quality data is required to feed into the ADS.

A real-time data warehouse can support many levels of increasing sophistication: reporting what happened, some form of analysis, providing prediction capabilities, operationalization, and ultimately becoming capable of making events happen.

D. DATA, TEXT AND WEB MINING

1. Data Mining Concepts and Applications

Today's organizations utilize analytical decision making to increase the speed and improve the quality of decision making. They use analytics to better understand their customers and to optimize the supply chains to maximize their returns on investment. The creation of huge databases introduced the idea of analyzing stored data to discover useful knowledge. As discussed earlier, data mining is a multidisciplinary field, the aim of which is discovering knowledge in large databases. Data mining utilizes statistical, mathematical, artificial intelligence, and machine learning methods to identify potentially useful patterns, and relationship in the data.

Data mining tools locate pattern in data using one of the following methods:

- Simple exploration models and tools such as SQL based queries, OLAP, and guided by human judgment,

- Intermediate models such as regression or decision tree or clustering,
- Complex models using neural networks, decision trees, and rule induction.

For instance, the supervised induction or classification is performed to analyze the historical data in order to generate a behavior prediction model.

Broadly, data mining can help foretell customer needs, monitor vehicle accidents and driver distractions, identify customer behavior, customizing medicine, or even mine financial transaction data to uncover terrorist funding.

2. Data Mining Techniques and Tools:

"Data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information—information that can be used to increase revenue, cut costs, or both" [11]. It focuses on numerous tasks such as rating clients by their likelihood to respond to an offer, estimating illness re-occurrence or hospital re-admission probability, identifying cross-selling opportunities, detecting fraud and abuse, and optimizing the parameters of a production line operation. Several data mining methods are developed and implemented in commercial software. The major algorithm used for data mining tools and techniques are: statistical methods, decision tree, case based reasoning, neural computing, intelligent agents, and generic algorithms.

Classification is the most common data mining technique. It analyses historical data to generate a model that would predict future behavior. The same model can then be used to predict unclassified data classes [2]. Neural networks and decision trees are good examples of techniques used for classification. Neural networks mimic the activities of the human brain for information processing. Neural Networks need to be trained using historical data before they can be applied to new data sets to classify them. A Decision tree is a hierarchical data classification scheme that classifies entities into particular classes based on the value of the attributes of these entities. A root note is followed by a hierarchy of nodes, each node is labeled with an if-then question. Arcs connect the nodes and cover all possible responses [2].

Clustering divides a database into groups of records with similar characteristics. It differs from classification in that the characteristics of the clusters are not known which requires an expert interpretation before the results can be used [2].

Association is based on discovering relationships between items occurring together. This technique is very useful for the retailer who can make a good interpretation of items that sell together [2]. Association analysis is also called Market Basket Analysis.

Sequence discovery is the discovery of associations over time [2]. This sequential technique allows, for example, identifying customer behavior over time, which can increase profits or eliminate fraud [2].

Linear and nonlinear regression are statistical techniques that allow estimating or predicting a numeric value, such as sales figures, based on a historical set of data [2]. Forecasting is also based on estimating future values using patterns in the data. Time-series methods are often used to predict future sales [2].

3. Text Mining

Text mining is the process of extracting useful patterns and relationships from large amounts of textual data. Text mining is similar to data mining in purpose and the processes used. While data mining looks for patterns in structured data, text mining is applied to unstructured data such as Word documents, PDF files, e-mails, XML files, and so on. Data mining, however, complements text mining. Text mining can be thought of a two-step process. The first imposes structure on structured data, followed by extracting potentially useful patterns and relationships from the now structured text-based data using data mining techniques.

Text mining is useful in application where large amount of textual data is being generated or collected, such as law, finance, medicine, intelligence, news gathering, technology and marketing. The most important applications of text mining include information extraction, topic tracking, summarization, categorization, clustering, concept linking, and question answering [2].

4. Web Mining

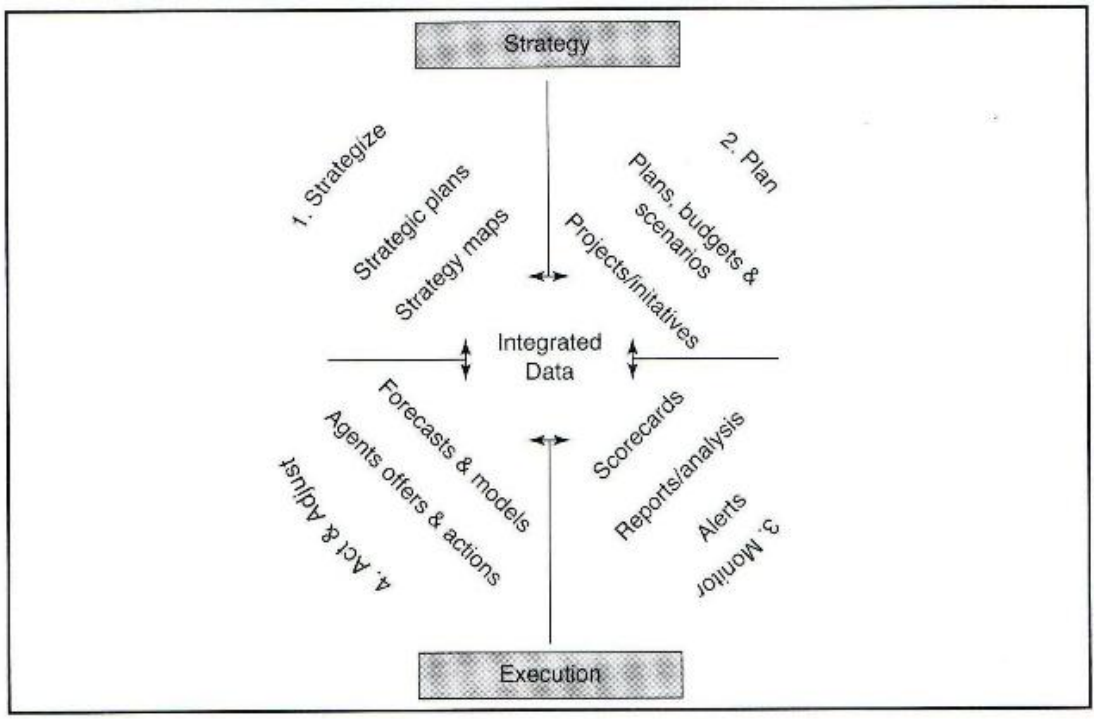
Web mining is defined as discerning and analyzing valuable and useful information from the web using mining tools. Web mining goes beyond mining the textual data in the form of web pages to include mining hyperlink information and usage information available in web logs, all of which provide rich data for knowledge discovery. Thus, web mining consists of web content mining, web structure mining, and web usage mining.

E. BUSINESS PERFORMANCE MANAGEMENT

1. Overview

Business Performance Management (BMP) can serve as “A real-time system that alert managers to potential opportunities, impending problems, and threats, and then empowers them to react through models and collaboration” [10]. BPM is an outgrowth of BI and incorporates many of its technologies, applications, and techniques. It includes a set of closed-loop processes between strategy and execution in order to optimize business performance which is achieved by: setting goals and objectives, establishing initiatives and plans to achieve those goals, monitoring actual performance against the goals and objectives, and taking corrective action.

In the following sections, we discuss these processes in some detail.



Source: W. Eckerson. *Performance Dashboard*, Wiley. Hoboken, NJ, 2006.

Figure 7 FBPM closed-loop process. (From [2])

2. Strategize: Where do You Want to Go?

Strategy deals with the ultimate question "Where do we want to go in the future?" The answer to this question is contained in a strategic plan, which is very similar to a map that guides from a current state to a future one.

In order to produce a strategic plan, a number of tasks must be accomplished. The first task is to conduct a situation analysis to review the organization's current situation and establish a baseline. The second task is to determine a planning horizon, which is usually 3 to 5 years, depending on numerous organizational factors. The third and fourth tasks conduct an environment scan and identify critical success factor (CSF); those factors that define the things that an organization must excel at to be successful. Next, the fifth task completes a gap analysis between where the company is and where it would like to be. This is followed by a sixth task that creates a strategic vision or a mental image of what the organization should look like in the future. The seventh task develops a business strategy based on the data from the previous steps. Finally, the eighth task

identifies strategic objectives–broad statements or general course of action prescribing targeted directions for an organization, and strategic goals—quantified objectives with a designated time period.

3. Plan: How do You Want to Go?

Once the "what" question is well defined, operational managers can specify how it will be implemented by developing an operational plan as well as a financial plan. An operational plan translates an organization's strategic objectives and goals into a set of well-defined tactics and initiatives, resources requirements, and expected results. It can be either tactic-centric to meet objectives established in a strategic plan, or budget-centric that sums the targeted financial values. A financial plan uses an organization's strategic objectives and key metrics as drivers for the allocation of an organization's tangible and intangible assets.

4. Monitor: How are You Doing?

In order to ensure that the organization is performing as expected, monitoring strategy must be established. A monitoring plan should address two issues: what to monitor and how to monitor. Many organizations use a diagnostic control system to monitor their performance and correct any deviations. As Figure 8 shows, a diagnostic control system is a cybernetic system that has inputs, a process, and benchmark against which to compare the outputs, as well as a feedback loop. In order for an information system to be a diagnostic control system, it must enable setting a goal in advance, allow to measure outputs, and calculate the performance variance that will be used as feedback to alter the input guiding performance to goals.

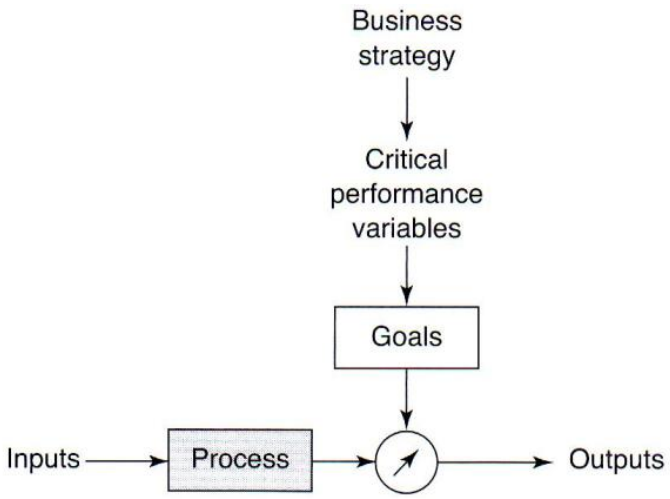


Figure 8 Diagnostic control system. (From [2])

5. Performance Measurement

According to Simons (2002), performance measurements systems help compare real results with strategic objectives by means of periodic feedbacks reports showing progress against goals [1]. Many current performance measurements system use some variant of the balanced scorecard (BSC). BSC methodology is a holistic vision of a measurement system tied to the strategic direction of the organization and based on a four-perspective view of the world: Financial measures supported by customer, internal, and learning and growth metrics Using financial data in performance measurement systems presents several limitations. First, financial data are provided by organizational structures not by the processes that produced them. Besides, these measures are only reporting what happened not what is likely to happen. In addition those measures deal with the short term rather than the long term.

A successful performance measurement system must be able to rapidly identify opportunities and problems, allocate resources after stating priorities, track any strategy change, define responsibilities and reward accomplishments, improve processes when the data affirm it, and promptly generate plan and prediction.

6. BPM architecture and Applications

BPM architecture is defined by its logical and physical design. As indicated in Figure 9. BPM consist of a database, an application, and user interface tiers. Data sources to a BPM can be provided by an enterprise resource planning (ERP) system or a data warehouse or external data such as market research data.

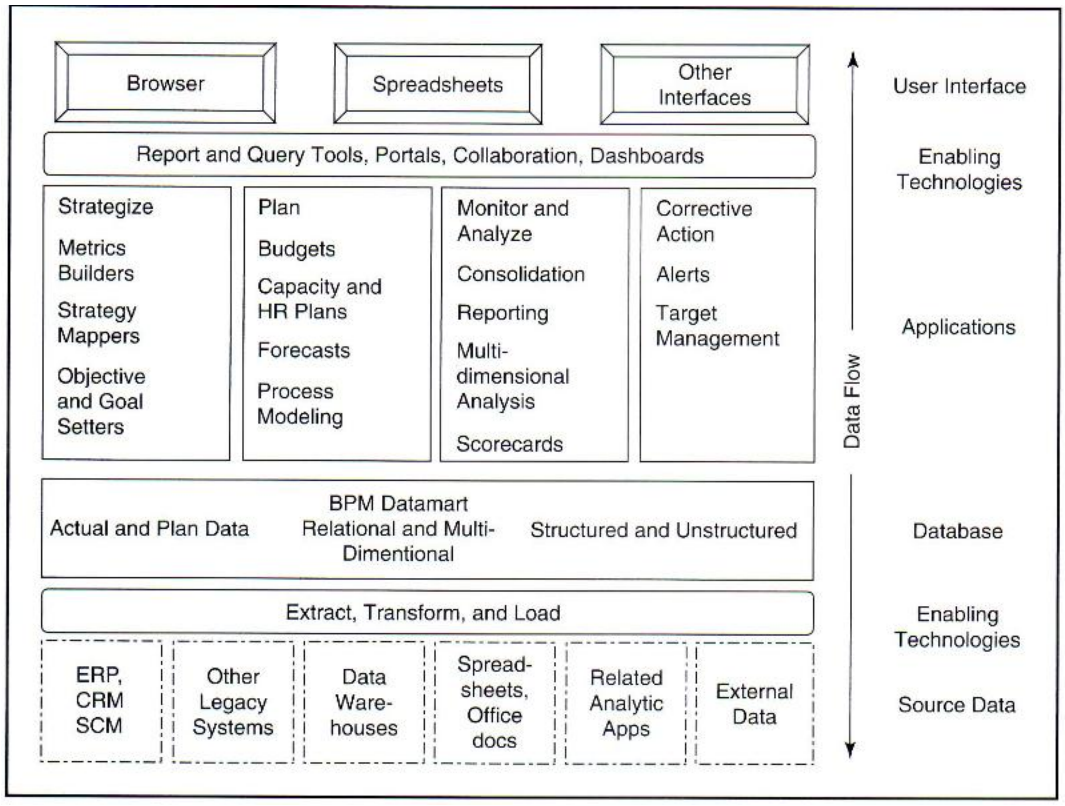


Figure 9 BPM architecture. (From [2])

BPM can be performed in budgeting, planning, forecasting, profitability modeling and optimization, scorecard applications, financial consolidation, and statutory and financial reporting.

7. Performance Scorecards and Dashboards

Dashboards and scorecards offer a visual picture of selected information in a single view so that information can be explored and digested easily by top executives.

The fundamental challenge of dashboard design is to display all the required information on a single screen, clearly and without distraction, in a manner that can be assimilated quickly [9].

Figure 10 Sample performance dashboard. (From [6])

Performance dashboards (Figure 10) are used to monitor operational performance. While performance scoreboards (Figure 11) are used to chart progress strategic goals. Dashboards offer tactical assistance. Scorecards describe the progress over time of some entities.

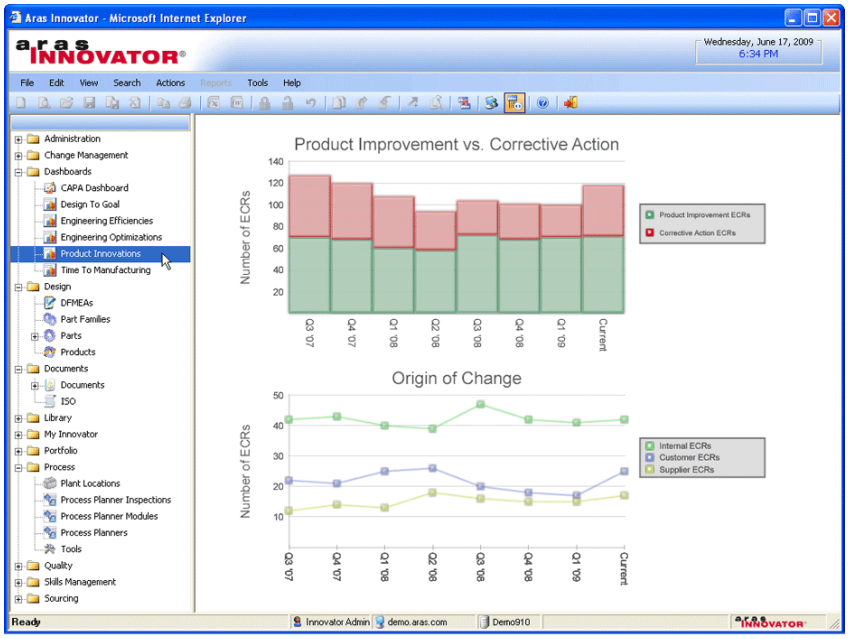


Figure 11 Performance scoreboards.

In the remaining chapters, we will analyze three state-of-the-art BI tools that accomplish many of the capabilities presented in this chapter.

III MEGAPUTER POLYANALYST DATA AND TEXT MINING SUITE

A. INTRODUCTION

This chapter describes Business Intelligence implementation in PolyAnalyst, a data and text mining tool developed by Megaputer. It describes its major data mining and machine learning as well as its querying and reporting capabilities. This chapter is organized as follows. Section A overviews PolyAnalyst and its main functionality. Section B presents Data mining algorithms. Section C describes the integration process. Section D deals with data manipulation in PolyAnalyst. Section E illustrates a set of data analysis algorithms. Finally, Section F describes a variety of visualization tools.

B. WHAT IS POLYANALYST?

PolyAnalyst is a commercial data mining software package from Megaputer. It uses a client-server architecture, where all processing takes place on a shared server. It provides the ability to read data from databases, statistical packages, HTML, Word and PDF files. An OLAP engine allows data to be gathered and "diced and sliced" prior to performing data mining algorithms. In addition, PolyAnalyst includes a variety of algorithms such as decision trees, fuzzy logic classification, genetic algorithms, neural networks, case-based reasoning, and text categorization.

There are two types of PolyAnalyst consumers: Data Analysts, and Business Users. Data Analysts perform data analysis scenarios using easy-to-use drag-and-drop interface and build reports to summarize the analysis result. Business Users interact with dynamic reports and executive dashboards that show key performance indicators in a comprehensive graphical display.

PolyAnalyst offers a multistrategy data mining suite including a set of Machine Learning (ML) algorithms for diverse mining tasks along with a structured data and text processing tools. It allows a deep integration especially when applying models to external databases through the OLE DB protocol, and exporting models to XML.

Figure 12 depicts PolyAnalyst architecture and main functionality.

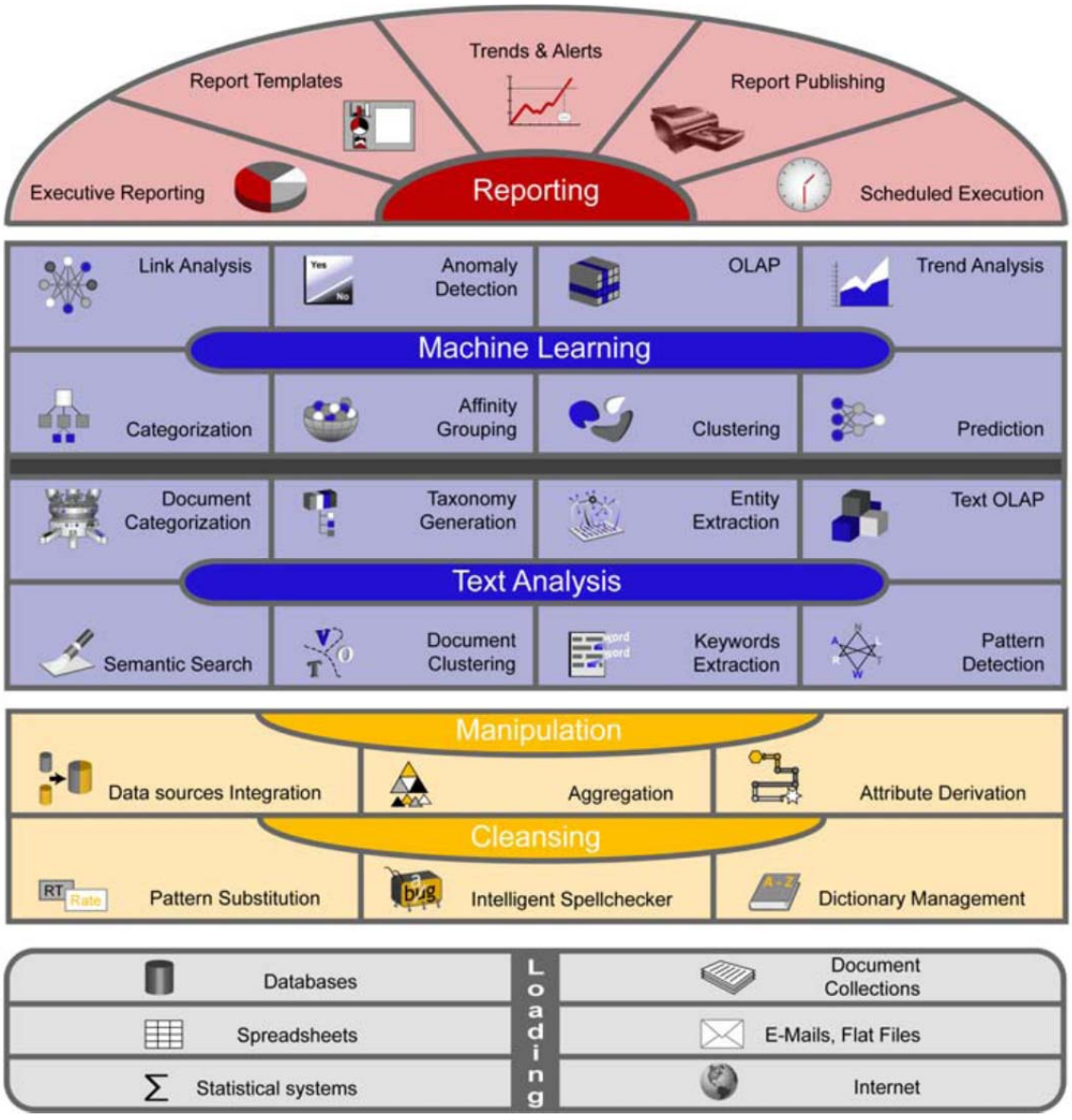


Figure 12 PolyAnalyst 6 features. (From [11])

C. DATA MINING IN POLYANALYST

In the business world, a competitive edge is usually obtained from knowledge. Knowledge can be extracted from existing data through the process of Data Mining.

Several definitions exist for data mining. The following two definitions capture the essence of data mining.

"Data mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner" [16].

"Data Mining is the process of finding new and potentially useful knowledge from data" [17].

Generally, data mining addresses various tasks such as rating clients by their likelihood to respond to an offer, identifying cross-selling opportunities, detecting fraud and abuse, estimating illness re-occurrence or hospital re-admission probability, optimizing the parameters of a production line operation, and predicting network peak loads.

Megaputer PolyAnalyst supports a variety of data mining algorithms such as Predictive Neural Networks, Classification Neural Networks, Rule Induction, Linear Regression, Logistic Regression, Case Based Reasoning, Bayesian Networks, CHAID, Decision Trees, R-Forests, Association Rule Learning, Temporal Association Learning, Anomaly Detection, Healthcare Fraud Signatures, Support Vector Machines, Naive-Bayes Classification, Expectation Maximization Clustering, Correlation Analysis, and Instance Based Reasoning. A summary description of the most important algorithms will be presented in Section E.

PolyAnalyst performs data pre-processing and modeling, as well as results reporting and delivery. PolyAnalyst supports the data mining tasks of predicting, affinity grouping, classification, clustering analysis, link and multi-dimensional analysis, patterns analysis, and interactive graphical reporting.

D. INTEGRATION PROCESS IN POLYANALYST

PolyAnalyst provides the capability to access data stored in commercial databases, some proprietary data format (such as Excel and SAS), as well as popular document formats. This capability is enabled through OLE DB protocol (Object Linking and Embedding for Database) or ODBC protocol (Open Database Connectivity). It also allows on-the-fly integration of data from disparate sources.

As shown in Figure 12, PolyAnalyst allows users to load data from databases, spreadsheets, statistical systems, document collections, e-mails, flat files, and the Internet.

PolyAnalyst offers the ability to integrate outputs into external application. PolyAnalyst provides several models to score data in different external databases through OLE DB protocol. In addition, it delivers models to external applications in the format they understand—XML. And integrates dedicated machine learning components in existing decision support systems

E. DATA MANIPULATION

Performing data analysis in PolyAnalyst consists of building analysis processes by defining a succession of work steps as a project flowchart such as the one shown in Figure 13. A process consists of a number of nodes that are connected to each other. A node processes an input to generate an output which could provide an input to another node. The action of the nodes is also called task, operation, or function. Nodes are grouped by their functions such as data source nodes which are used to import data into the analysis process from diverse data sources such as Microsoft Access, Microsoft Excel, or other data sources. Other node types include column operations, row operations, table operations, data analysis, text analysis, dimensional analysis and charts nodes.

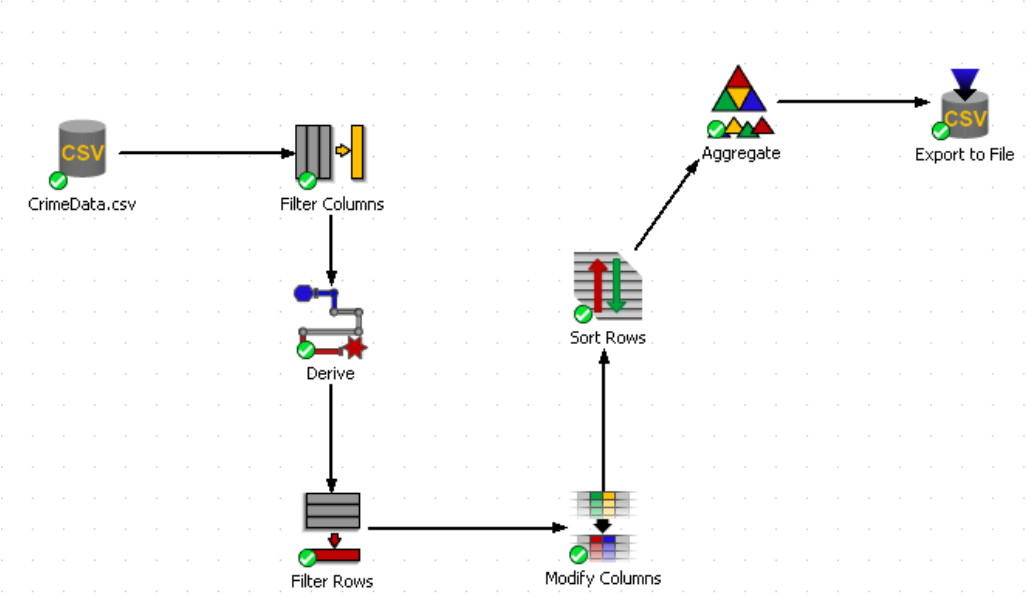


Figure 13 Project flowchart.

1. Dataset Statistics Viewing

PolyAnalyst offers the ability to toggle between dataset views and statistical information view for almost every dataset generated by a node [18]. As shown in Figure 14, a statistical view comprises the list of all columns names, the basic statistical properties for the selected column, and a histogram representing the distribution of values for the selected column. The example shown in Figure 15 represents a histogram of age distribution.

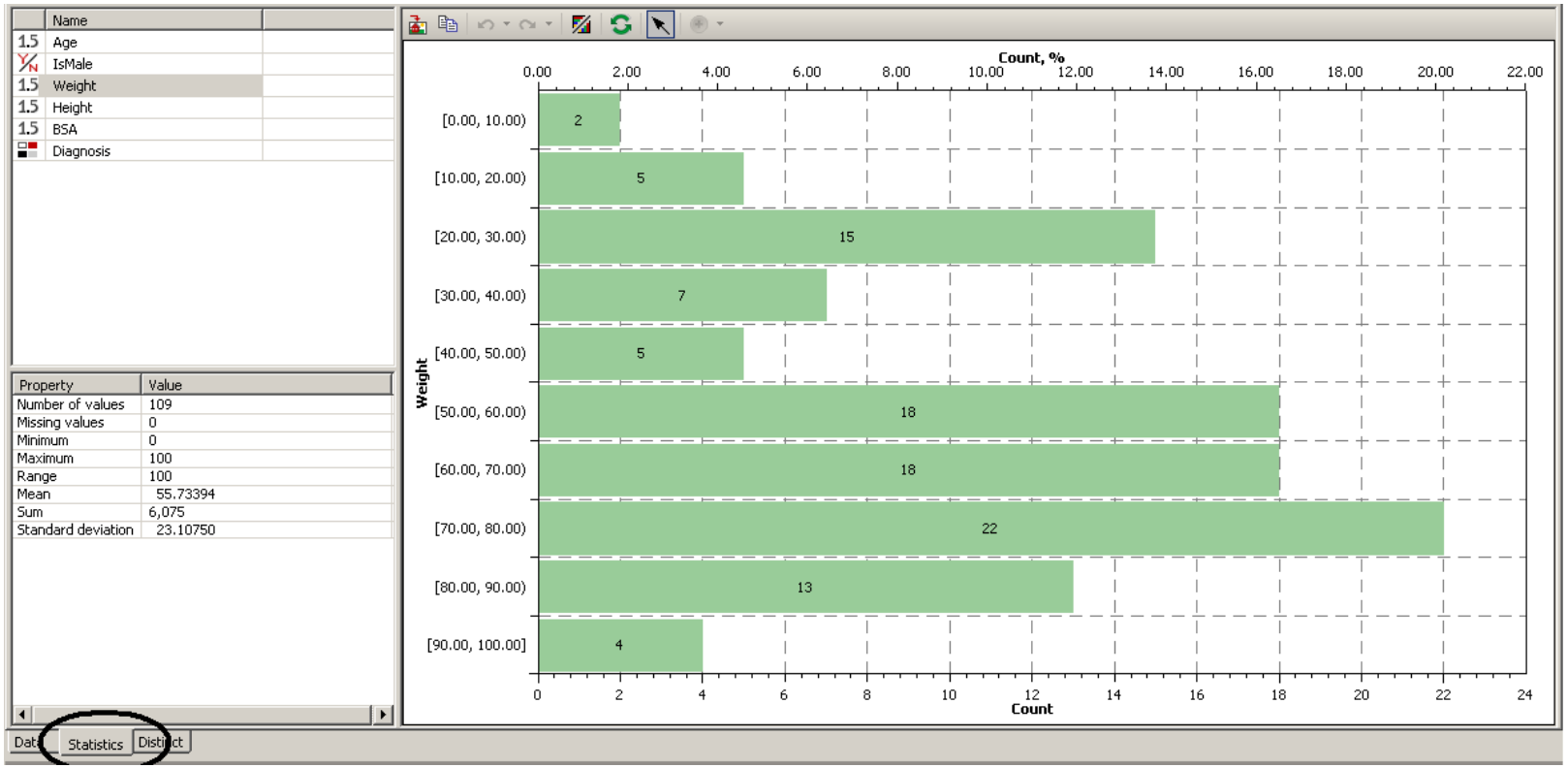


Figure 14 Displaying data in statistic view.

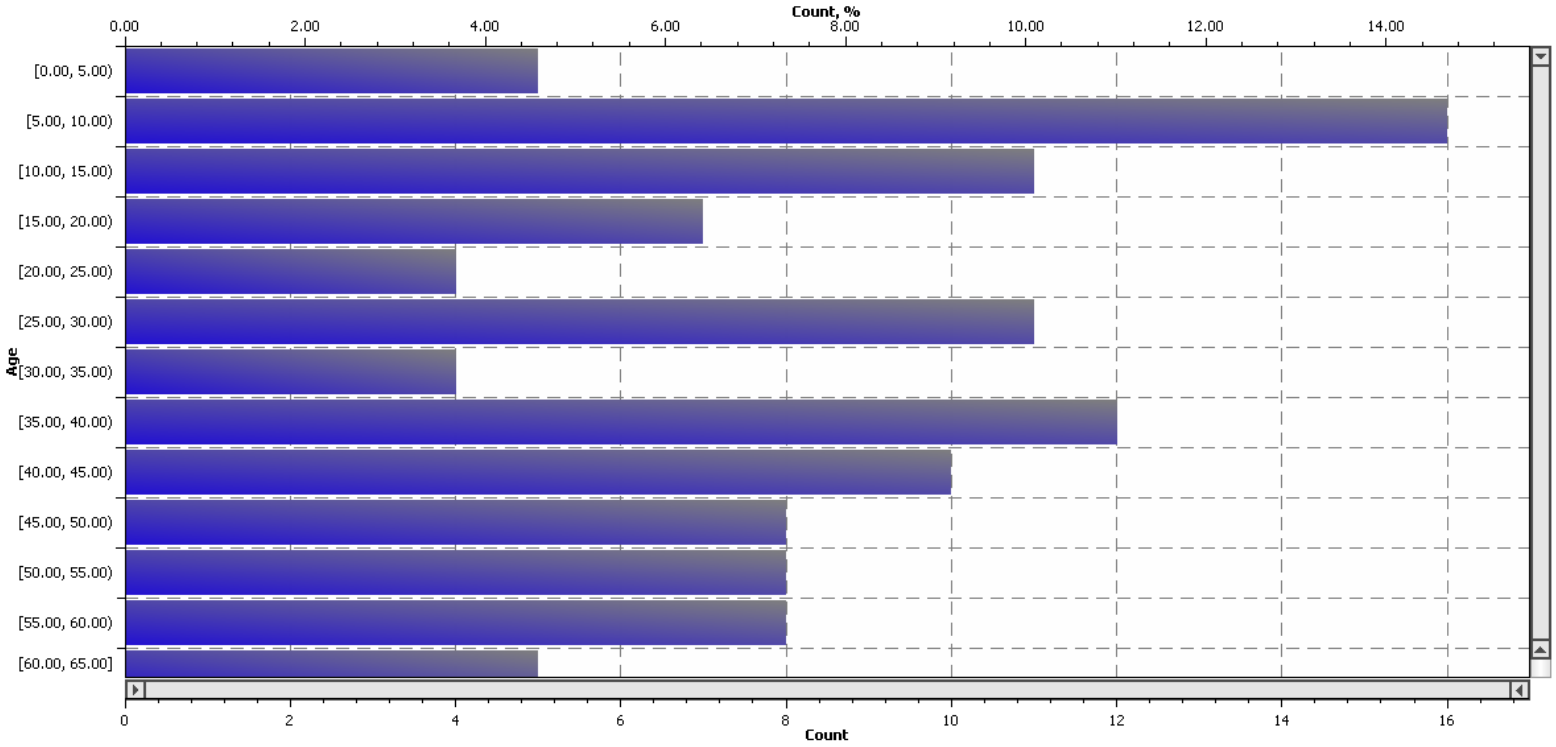


Figure 15 Age distribution.

2. Searching Data

PolyAnalyst offers the capability of storing millions of records and thousands of columns in datasets located at the PolyAnalyst Server [18]. PolyAnalyst Server performs data manipulation, querying and searching. In addition, the PolyAnalyst Analytical Client installed on the user computer can perform simple searches of data in the Data Viewer (less than 10,000 records). This raw client-side search does not utilize any pre-calculated index in contrast with searches occurring at the server which make use of the index.

3. Viewing Data

As shown in Figure 16, the data viewer tool allows displaying columns and rows of a dataset. The data viewer or data grid is a powerful module that is capable of displaying millions of records by loading only visible rows in the memory of the client computer and unloading old ones. This tool is common in many outputs in PolyAnalyst. The grid may be displayed in a report, as a result of drilling down on a chart, or as the output of an analytical node.

1.5 Age	IsMale	1.5 Weight	1.5 Height	1.5 BSA	Diagnosis
61.00	yes	65.00	172.00	1.77	MVD
0.00	yes	74.00	167.00	1.83	MVD
17.00	yes	86.00	178.00	2.04	TF
51.00	yes	71.00	174.00	1.85	MVD
26.00	no	53.00	168.00	1.59	ASD
39.00	no	45.00	164.00	1.46	ASD
20.00	no	56.00	163.00	1.60	ASD
9.00	yes	28.00	130.00	1.01	ASD
33.00	no	55.00	156.00	1.53	AVD
10.00	no	30.00	145.00	1.12	ASD
10.00	yes	30.00	149.00	1.15	AVD
12.00	yes	39.00	110.00	1.03	AVD
55.00	yes	94.00	170.00	2.05	IHD
48.00	yes	75.00	168.00	1.85	IHD
25.00	yes	73.00	182.00	1.94	ASD
45.00	no	69.00	170.00	1.80	ASD
8.00	no	25.00	130.00	0.96	AVD
6.00	yes	23.00	122.00	0.89	AVD
7.00	yes	20.00	116.00	0.81	ASD
40.00	no	60.00	164.00	1.65	MVD
4.00	no	17.00	110.00	0.72	AVD
6.00	no	20.00	126.00	0.86	TF
29.00	yes	58.00	164.00	1.63	ASD
11.00	no	32.00	148.00	1.17	ASD

Record 0 of 109

Data | Statistics | Distinct

Figure 16 Displaying data in a data grid.

While storing data in the data set, as shown in Figure 17, values can be formatted as a monetary value or a percentage before being processed in charts or algorithms.

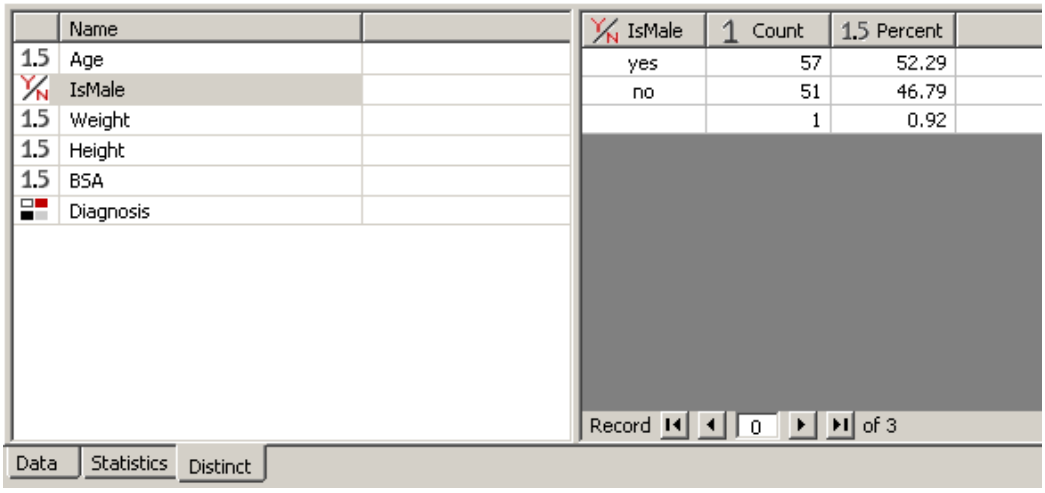


Figure 17 Displaying distinct values with count and percentage.

F. DATA ANALYSIS

Data analysis nodes perform analysis on input data and generate a model or report that can be used for viewing on screen, on the patterns generated by the analysis node from the input data. The input is typically a single table of data [18].

PolyAnalyst broadly classify its analysis nodes into four different categories: Structured data analysis, Text Analysis, Visualization, and Dimensional Analysis. Structured data analysis involves the development of statistical models based on structured data expressed in numbers, dates, and categories. Text analysis involves the analysis and modeling of textual data expressed in natural language [18]. Visualization involves the creation of charts or graphs that represent information about the data in some visual manner. Dimensional analysis nodes break apart a table of data according to user-created 'dimensions' or measures by which to 'slice and dice' data for understanding the data.

In the following sections, we present the main data analysis algorithms/nodes used in PolyAnalyst and provide a brief description of their functionality.

1. Find Laws

Find Laws is a nonlinear regression algorithm. This technique is utilized for predictive analysis. It is based on Megaputer's SKAT technology (Symbolic knowledge acquisition technology). The model is generated in SRL[1] expression and can be also scored along with other data using the Score node. Find Laws offers the capability to model relationships hidden in data, provides clearly discovered knowledge, and find all possible hypotheses.

2. Cluster (Localization of Anomalies)

Clustering or grouping similar records involves the theory of similarity. The objective of clustering is to maximize the resemblance between records from the same group and maximize the dissimilarity between records from different groups. This algorithm chooses best variables for clustering and groups the clusters of similar records in new dataset for further analysis. Cluster technique does not measure distances between points, but analyses variables distributions in hypercubes.

3. Find Dependencies (n-Dimensional Distributions)

Find Dependencies is an evaluation of the relationship between one variable and one or more other variables. When manipulating independent variables, the behavior of the dependent variable—also called the response variable or the target variable—is measurable using the Find Dependencies. This algorithm is used as preprocessing for Find Laws (FL). The Find Dependencies node defines most influential variables, detects multi-dimensional dependencies, and predicts the response variable in a table.

4. Classify (Fuzzy Logic Modeling)

The Classify algorithm in PolyAnalyst is a fuzzy-logic based classification algorithm. It allocates data to one or two classes and provides the classification rule. This technique assists in finding the proximity degree between classes in order to perform a

[1] PolyAnalyst's Symbolic Rule Language (SRL) contains functions and operations for manipulating data and calculating results. SRL is an algebraic and flexible syntax relied upon by several nodes for logical processing of data.

better categorization of records of interest. This algorithm is used either by Find Laws, PolyNet Predictor, or LR.

5. Decision Ttree

A decision tree is an algorithm that represents a decision problem graphically in a form of a tree, as shown in Figure 18, including the possible outcomes of decisions made at each stage. A decision tree is used in different decision problems situations in various fields (security, Finance and Insurance, medicine, etc.).

A Decision Tree is used to classify cases to selected categories. It is based on information gain partition criteria and offers the ability to scale linearly with increasing number of records.

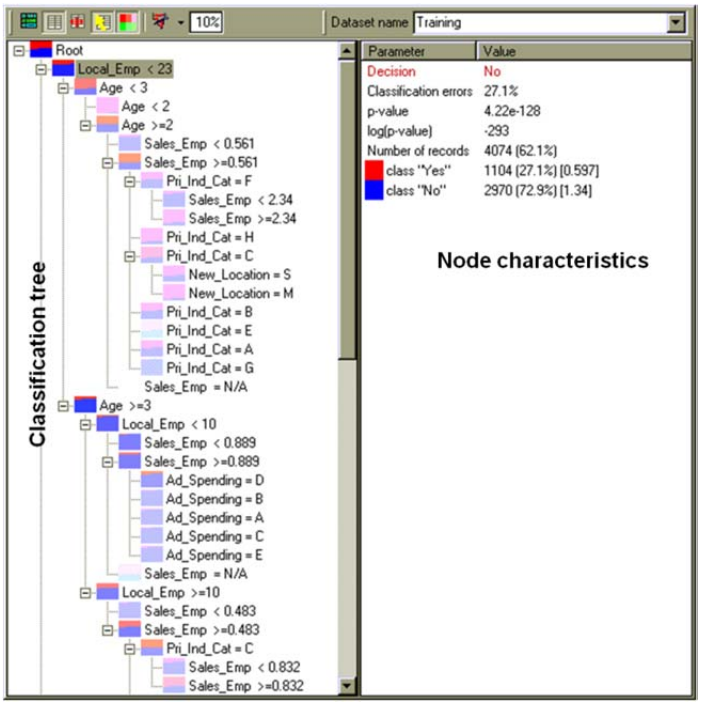


Figure 18 Decision Tree example.

6. PolyNet Predictor (GMDH-Neural Net Hybrid)

PolyNet Predictor is an algorithm that allows prediction of values of continuous attributes. It utilizes Hybrid GMDH-Neural Network method. It can work with large amounts of data and allows building the network architecture automatically.

7. Market Basket Analysis (Association Rules)

Market basket analysis is a data mining technique aimed at finding groups of items that often occur together in a transaction. This algorithm was originally used in retailing, yet it has been successfully applied in other domains.

8. Memory Based Reasoning (k-NN + GA)

Memory based reasoning is a data mining technique aimed at finding a collection of the most similar data and then forecasting the membership of new unknown data in the defined groupings. Memory based reasoning is mostly implemented using the k-Nearest Neighbor algorithm.

9. Linear Regression (Stepwise and Rule-Enriched)

Linear Regression is a statistical techniques used for prediction. This technique aimed at fitting a line through a set of points by minimizing the sum of the squares of the distance between the line and each data point. Stepwise linear regression in PolyAnalyst can work with unlimited number of attributes, and identifies the attributes that leads to the best linear prediction rule.

10. Discriminate (Unsupervised Classification)

Discriminate algorithm or unsupervised classification verifies what features of a selected data set discern it from the rest of the data. This algorithm does not require any target variable. It can be powered by PolyNet Predictor (PN), Linear Regression (LR), and Find Laws (FL).

11. Link Analysis (Visual Correlation Analysis)

Link analysis is aimed at identifying correlations among classes of categorical variables and displaying the results in a graph. Link analysis allows trends analysis, discovery of patterns of interest, as well as identifying correlation between variables.

12. Text Mining (Semantic Text Analysis)

Text mining in PolyAnalyst is performed via Text Analysis nodes. This approach is applied to text type columns in a table. Example text mining operations performed with this node include Phrase Extraction, Linear Classification, Link Terms, Spell Check, Keyword Extraction, Text Clustering, Search Query, and Entity Extraction. The Result of this process is a report of the extracted information from the analyzed texts.

G. REPORTING AND VISUALIZATION

This section describes a variety of visualization tools available in PolyAnalyst. Each of the visualization supported by PolyAnalyst can be integrated in a report.

1. Histograms

As shown in Figure 19, a histogram is a two-dimensional vertical chart utilized to better illustrate data distribution. Histograms can be added to reports using the Report Designer tool.

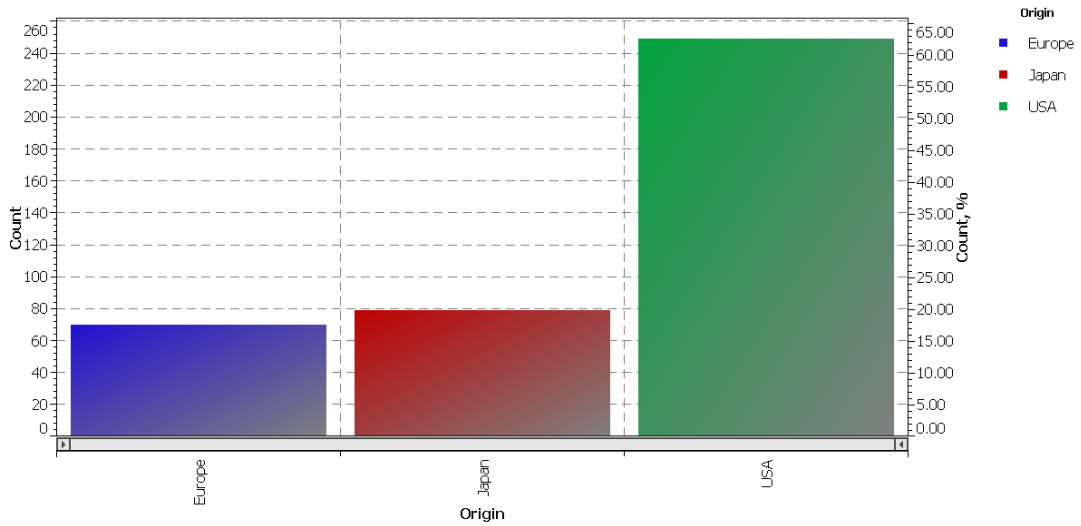


Figure 19 Histogram showing cars distribution by origin.

2. Line and Scatter Plots with Zoom and Drill-Through Capabilities

A line chart displays the relationship between two or more continuous variables where the data points are connected by lines. Figure 20 represents a scatter plot that displays the relationship between two or more continuous variables, but the data points are not connected by lines. A line/scatter plot chart helps visualize whether a positive, negative or no relationship exists between variables. A positive relationship is considered when an increase in one variable causes an increase in the other. Similarly, a negative relationship confirms that a decrease in the value of one variable leads to a decrease in the other. Whereas the absence of a relationship is illustrated when the plot is dissipated and no line or trend is shown on the plot. A linear predictive model is characterized by a straight line; while a curved line indicates a non-linear model. When no trend is suggested in a plot, normalizing, modifying the dimensions, or working on subsets is required to develop a predictive model.

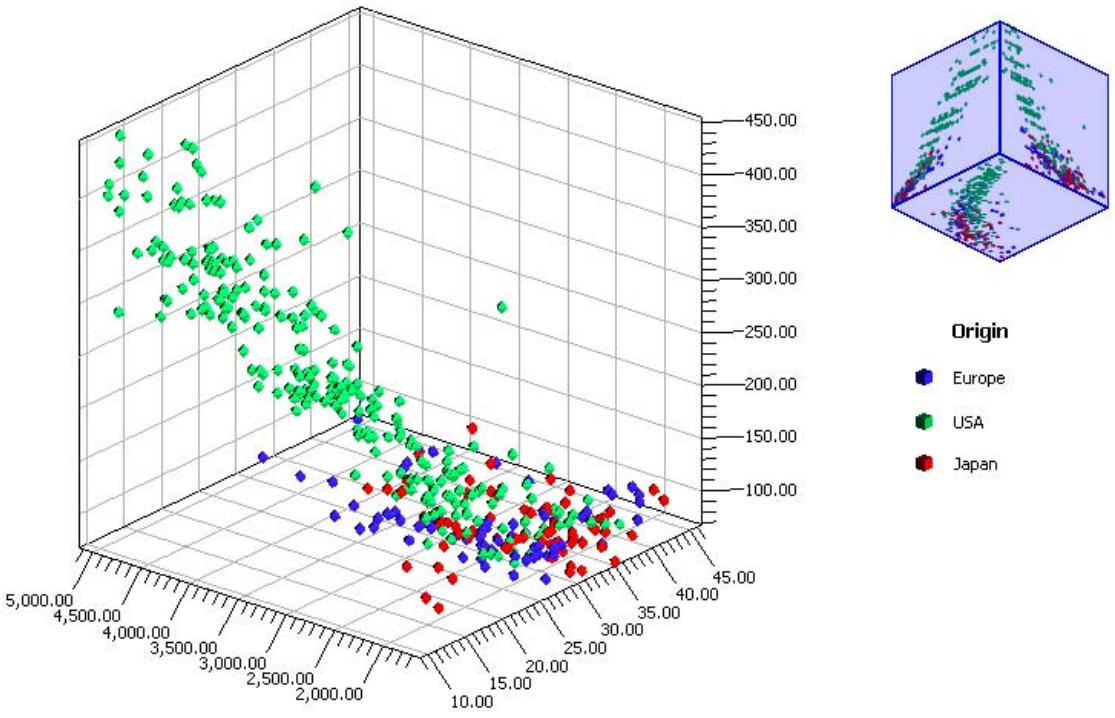


Figure 20 3D scatter plot.

PolyAnalyst's line chart displays data points produced by functions in a 2D coordinate plot. As shown in Figure 21, this functionality illustrates trends or to elucidate the variance of a set of attributes when another attribute changes. It allows choosing multiple Y axes on a single X axis.

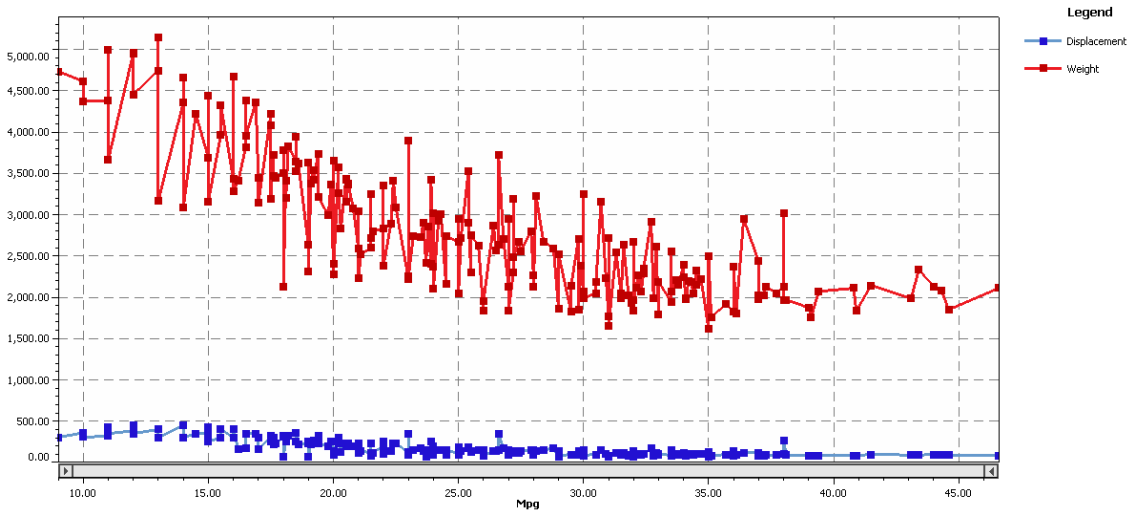


Figure 21 Line chart applied on Cars dataset.

3. Snake Charts

A Snake chart is a chart that shows the distribution level of a set of variables. It depicts the variation of categorical variable values with respect to the overall mean. This is useful to visualize high or low distributions across several variables at once.

As shown in Figure 22, a snake chart compares the variation of the variable's distribution as split according to the distinct values of a categorical variable. It also provides a comparison of high or low distributions of several variables at once.

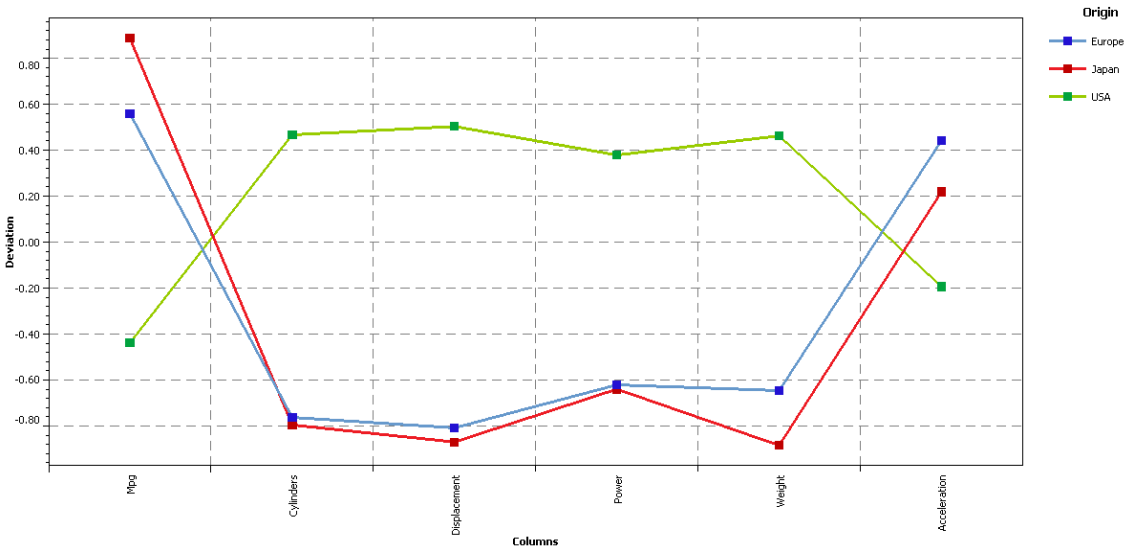


Figure 22 Snake chart display.

4. Interactive Charts

This chart type differentiates two types of pie charts: pie and doughnut charts. These charts are available in 2D flat charts, 3D dimensional charts, and 2D dimensional charts with slightly spread out slices (see Figure 9). Users can interact with these charts using the mouse and control buttons. They can resize, rotate, and manipulate the charts to further their understanding of the data. Furthermore, these charts allow user to drill down by selecting different areas of the chart.

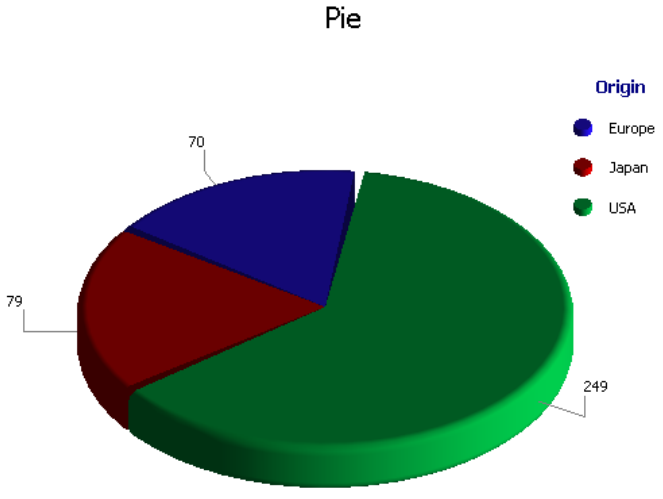


Figure 23 3D pie charts.

5. Lift and Gain Charts for Marketing Applications

The lift chart is useful in determining the rate of response in a direct-marketing campaign. The x-axis represents the percentage of all prospects contacted by a marketing campaign; the y-axis is the percentage of responders (those who, if contacted, would respond to the offer) that were reached by the campaign [18]. The gain chart is a graph that allows the visualization of data mining results through database marketing framework. It uses the specific business setting with the significance of data mining model as produced by a lift chart in order to determine the maximum profit conditions. Figure 24 illustrates the response to prospects using these charts for marketing applications.

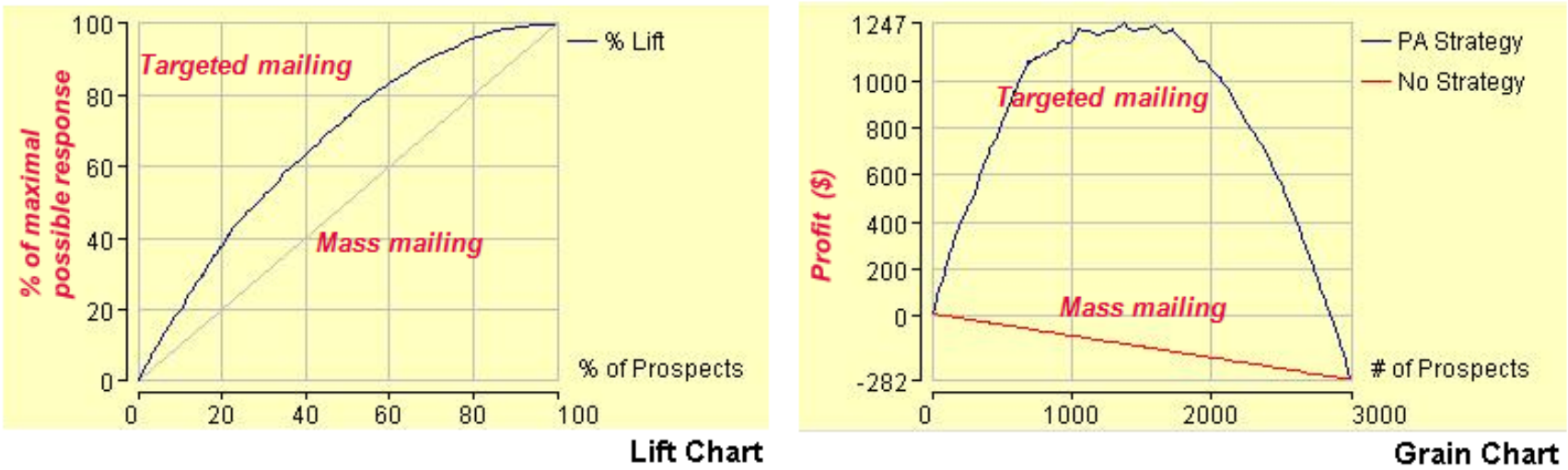


Figure 24 Lift chart versus grain chart for marketing campaign.

Megaputer PolyAnalyst is powerful software for data mining, text mining, and web mining. It supplies the analyst with several capabilities to discover unseen relationships in data and identify patterns and relationships, thus enabling business intelligence. The discovered patterns and relationships allows for faster and better decision making.

THIS PAGE INTENTIONALLY LEFT BLANK

IV. ORACLE BUSINESS INTELLIGENCE TOOLS

This chapter describes Oracle Business Intelligence (OBI), a high level querying and reporting tool. Oracle Business Intelligence Enterprise Edition is a complete business intelligence platform that includes a full set of analytics, OLAP, reporting as well as scorecards. The chapter is organized as follows. Section A describes BI answers, an interactive reporting and charting component of OBI. Section B describes the OBI Interactive Dashboard, and its main components. Section C discusses BI Delivers, and OBI component that allows business activity monitoring and alerting through multiple channels. Section D overviews the segmentation and list generation components of OBI. Section E presents the disconnected analytics of OBI intended to support mobile users who are usually disconnected from the Corporate Network. Section F illustrates Oracle Publisher and the enterprise reporting and distribution tool. Section G presents the interactive reporting component BI Publisher, and finally, Sections H, I, and J wrap up the chapter with a discussion of the SQR Production Reporting, Financial Reporting, and Web Analysis, respectively.

A. BI ANSWERS

In the OBIEE environment, users do not have to manipulate complex database structures. Instead, users are presented with a logical view of the information that they can manipulate. BI answers is the component of the OBI that consists of interactive charts, pivot tables and reports that can be easily manipulated by the business users at the logical view level. It represents a new generation of Ad-hoc Reporting and Querying tool. The following figure shows an example of BI Answers application.

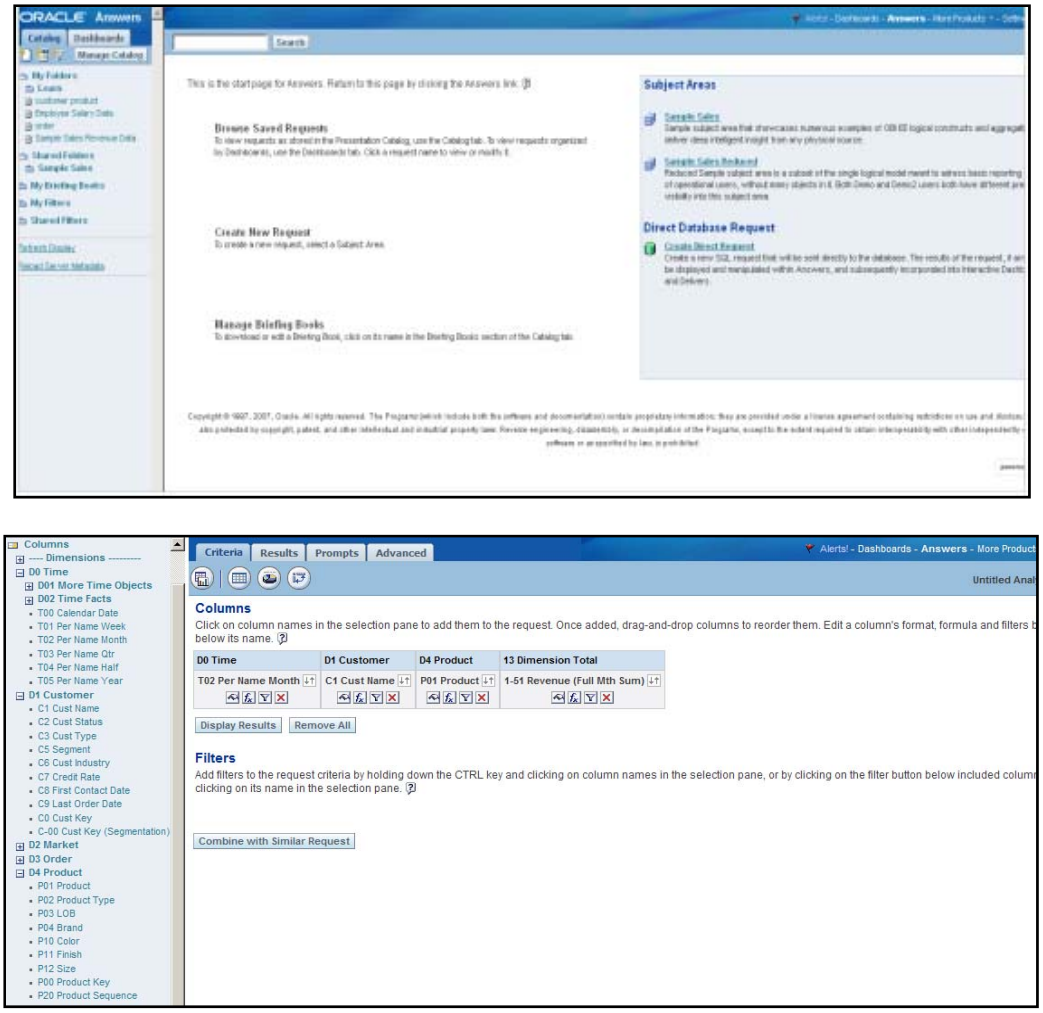
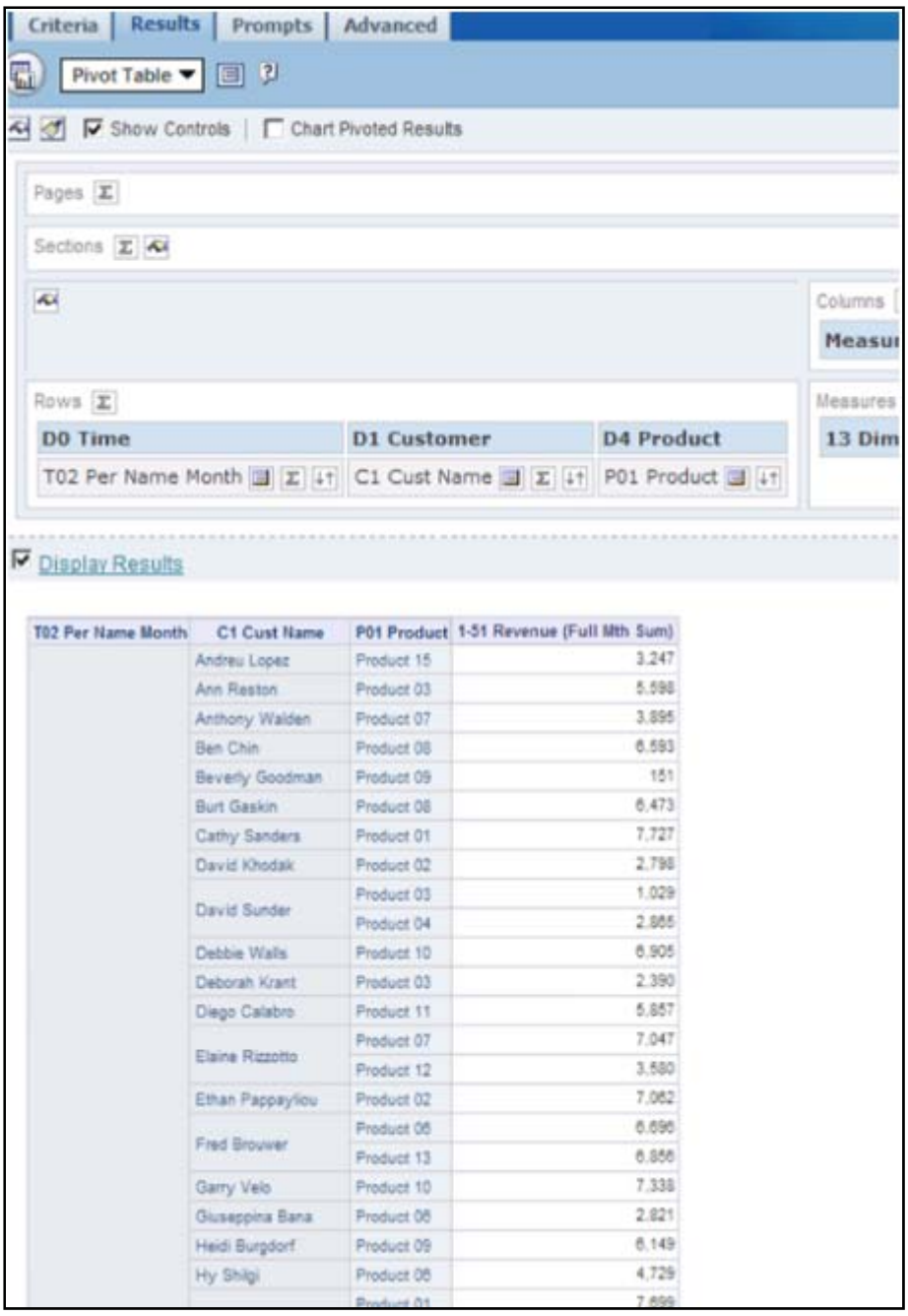


Figure 25 Answers criteria selection.

OBIEE installation comes with an Oracle Virtual Machine template. The template includes a data source that communicates to an OLTP backend. “Sample Sales” is a BI repository included in the installed Oracle Virtual Machine template. Figure 25 shows how to build the criteria within BI answers. Time, Customer, and Product are the dimensions chosen from “Columns.” Once the criteria are built, results may be displayed as a table, pivot table, or charts.

Figure 26 Answers result displayed as a table.

Figure 26, displays the revenue for each customer for each product per month. The result is displayed and grouped by Name Month then Customer Name which makes the interpretation of the decision maker more focused.



T02 Per Name Month	C1 Cust Name	P01 Product	1-51 Revenue (Full Mth Sum)
	Andreu Lopez	Product 15	3,247
	Ann Reston	Product 03	5,598
	Anthony Walden	Product 07	3,895
	Ben Chin	Product 08	6,593
	Beverly Goodman	Product 09	151
	Burt Gaskin	Product 08	6,473
	Cathy Sanders	Product 01	7,727
	David Khodak	Product 02	2,798
	David Sunder	Product 03	1,029
		Product 04	2,865
	Debbie Walls	Product 10	6,905
	Deborah Krant	Product 03	2,390
	Diego Calabro	Product 11	5,857
	Elaine Rizzotto	Product 07	7,047
		Product 12	3,580
	Ethan Pappaylou	Product 02	7,062
	Fred Brouwer	Product 06	6,696
		Product 13	6,856
	Garry Velo	Product 10	7,338
	Giuseppina Bana	Product 06	2,821
	Heidi Burgdorf	Product 09	6,149
	Hy Shilgi	Product 06	4,729
		Product 01	7,699

Figure 27 Answers result displayed as a pivot table.

Figure 27 displays the revenue for each customer for each product per month in a pivot table.

Depending on the application, results in OBIEE can be displayed in different

formats. Figure 28 illustrates the use of interactive graphs to display the results, while Figure 29 presents the results of a user query as a map overlay with drill down capabilities.

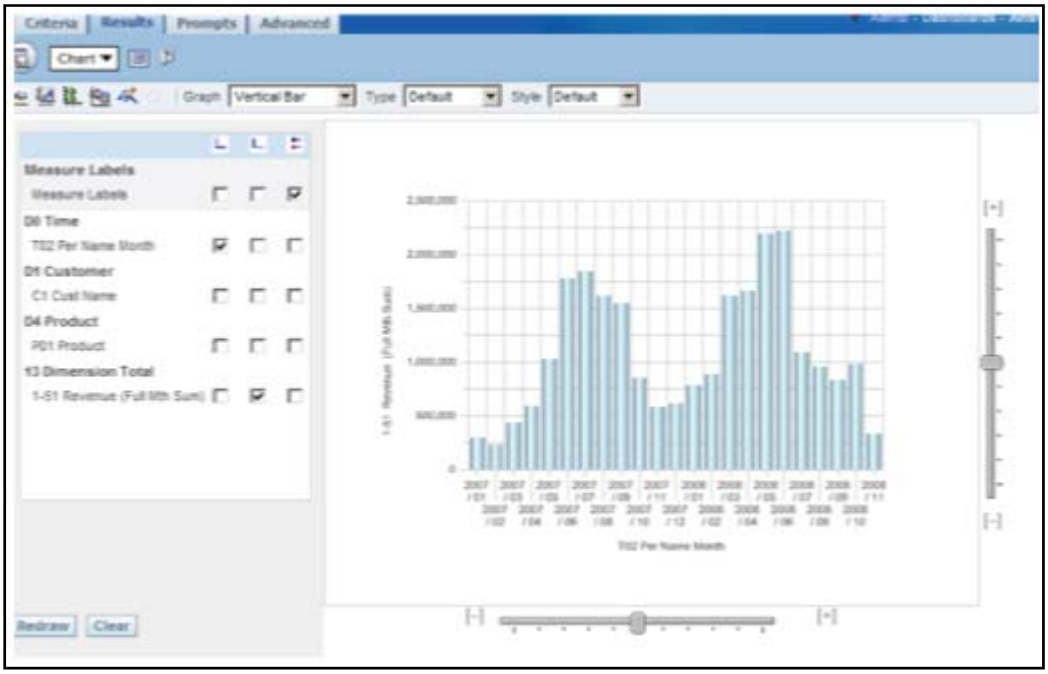


Figure 28 Answers result displayed as a graph.

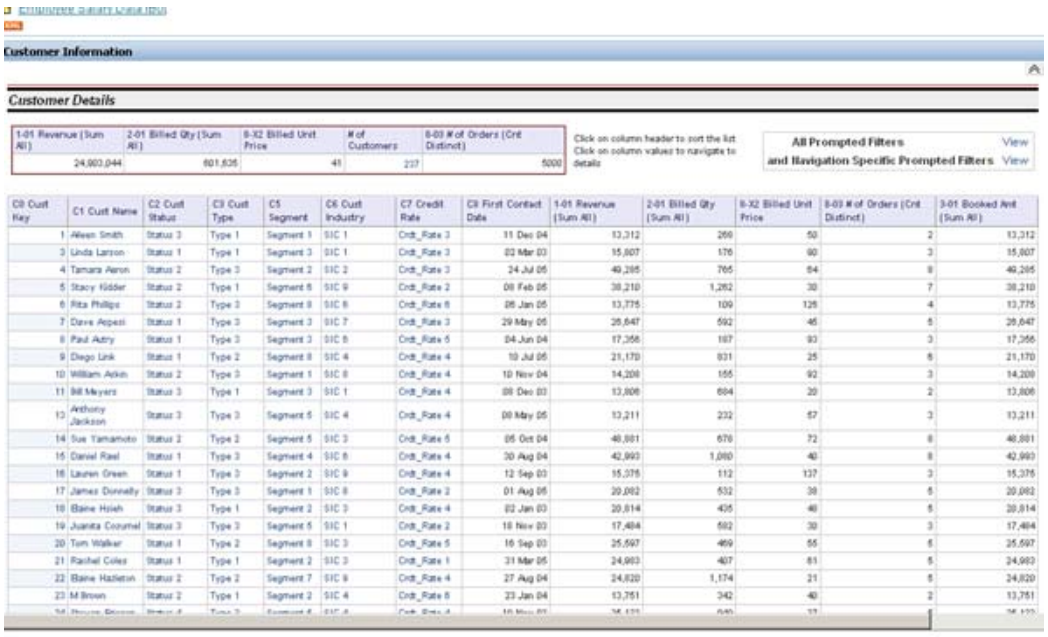
Figure 29 Answers results displayed as a map Overlay. (From [19])

The components generated by answers like tables, graphs, views, are very useful for the decision maker. Answers component are then published in Oracle BI Interactive Dashboards to be discussed in the next section.

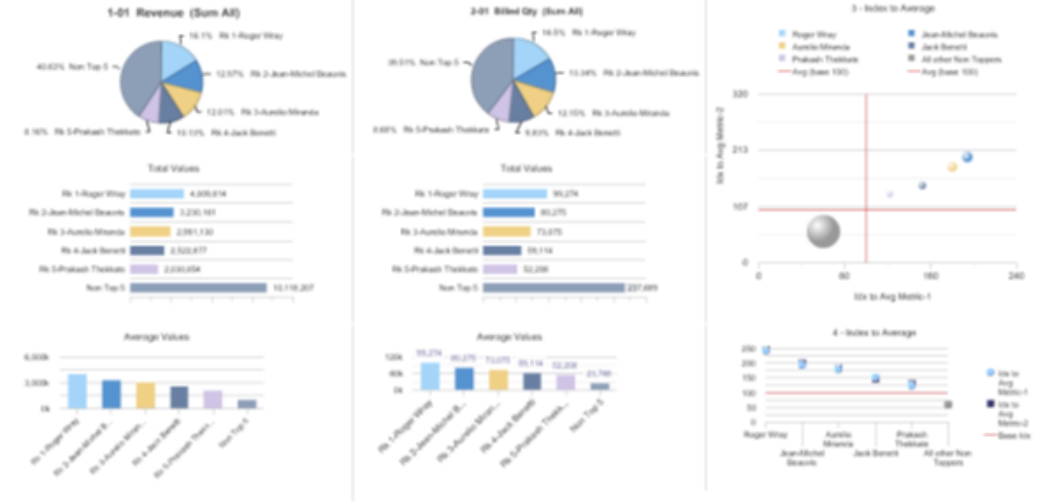
B. BI INTERACTIVE DASHBOARD

BI interactive dashboard is integrated with BI answers. Based on their roles in the organization, business users have the ability to personalize an appropriate dashboard. This dashboard includes an interactive collection of content and applications. Figure 30 illustrates an example of a personalized dashboard that includes various tables and graphs, discussed in the previous section, to provide the decision maker with a well summarized view of his area of interest, thus supporting his decision making process.

Oracle BI Dashboard is a presentation layer tool integrated with BI Answers which allows users to perform Ad-hoc analysis and report against the business model.



Multi Metrics Proportional Top Ns



Toppers	Count	Rk	Sum Metric-1	Avg Metric-1	Idx to Avg Metric-1	Rk	Sum Metric-2	Avg Metric-2	Idx to Avg Metric-2
Roger Wray	1	Rk 1-Roger Wray	4,000,014	4,000,014	242	Rk 1-Roger Wray	66,274	66,274	240
Jean-Michel Beauvis	1	Rk 2-Jean-Michel Beauvis	3,230,161	3,230,161	195	Rk 2-Jean-Michel Beauvis	60,275	60,275	200
Aurelio Miranda	1	Rk 3-Aurelio Miranda	2,991,130	2,991,130	180	Rk 3-Aurelio Miranda	73,075	73,075	182
Jack Benetti	1	Rk 4-Jack Benetti	2,522,677	2,522,677	152	Rk 4-Jack Benetti	56,114	56,114	147
Prakash Thakkate	1	Rk 5-Prakash Thakkate	2,030,954	2,030,954	122	Rk 5-Prakash Thakkate	52,209	52,209	130
All other Non Toppers	10	Non Top 5	10,110,207	1,011,021	61	Non Top 5	237,699	23,769	59
Grand Total	15	Non Top 5	24,900,044	1,660,203	100	Non Top 5	601,635	40,109	100

History

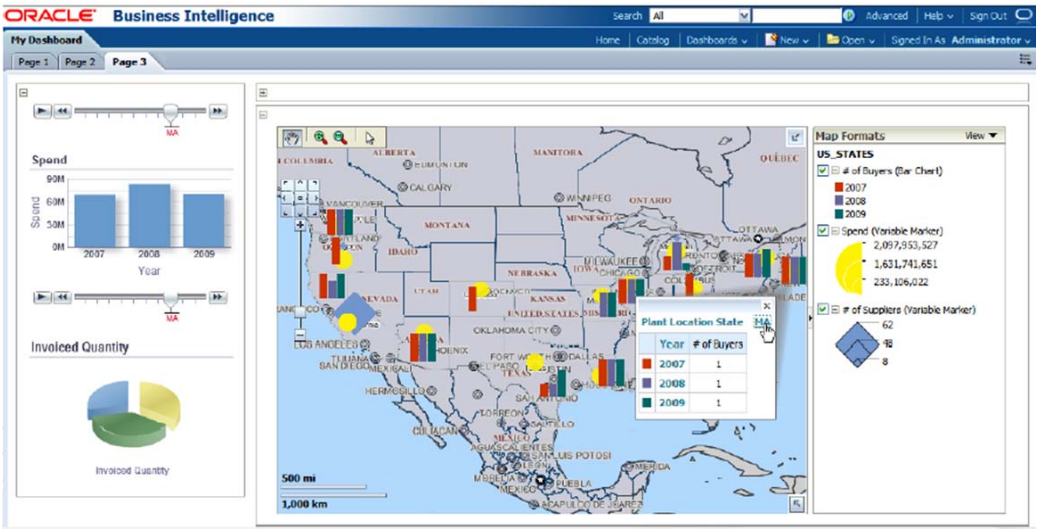


Figure 30 Interactive Dashboard. (After [19])

C. BI DELIVERS

BI delivers is a component of OBI that allows business activity monitoring and alerting through multiple channels such as e-mail, dashboards and mobiles. Figure 7 is the main screen of showing user options for running the application.

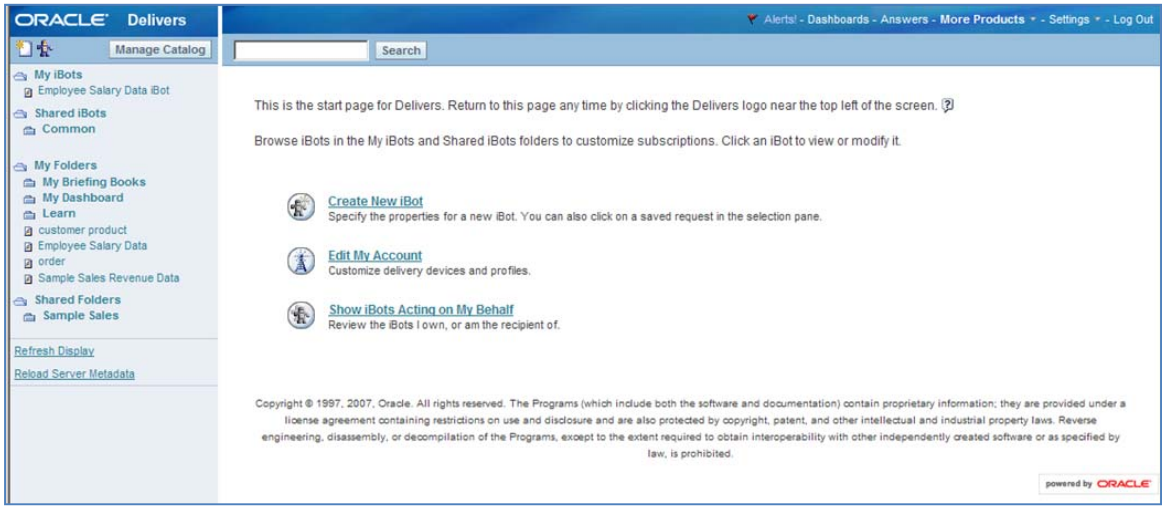


Figure 31 Oracle Delivers main screen.

D. SEGMENTATION AND LIST GENERATION

This component allows creating a class of customers or prospects on which the studies or subject areas such as campaign history, orders, or service will be focused. The Segmentation Designer provides the number of customers that respond to the criteria. Every column added to the criteria induces a change in the segment. Since the criteria are reevaluated against the current values in the database, a segment may change over the time. Using the Manage Marketing Jobs link, the administrator can display a jobs and cache entries.

E. DISCONNECTED ANALYTICS

Oracle BI Disconnected Analytics enables full analytical functionality for the mobile professional who is disconnected from the Corporate Network. It allows disconnected users to view analytics data, Oracle BI Interactive Dashboards, and queries. Typically, mobile users connect their personal computers to an OBIEE server and download an Oracle BI Disconnected Analytics application. Subsequently, they can disconnect their machines from the network and still able to view dashboards and queries.

OBIEE offers two disconnected solutions: Oracle BI Briefing Books and the Managed Oracle BI Disconnected Analytics. A brief explanation of each follows.

1. Oracle BI Briefing Books

Briefing Books allow disconnected users who are working offline to put a static snapshot of Oracle Business Intelligence content on their machines. This approach allows users to use the static content in Oracle BI Briefing Books to fulfill such tasks like managed reporting and lightweight content distribution. Oracle BI Briefing Books can be scheduled and delivered using Intelligent Bursting and Output Tool (iBot), which is also used for alerts and scheduled reports over any web enabled device.

2. Managed Oracle BI Disconnected Analytics

Managed Oracle BI Disconnected Analytics is centrally administered and managed. Once their local databases are populated, disconnected users can connect to a local dashboard and view similar Interactive Dashboards to the one installed in the Oracle Business Intelligence server. This tool offers most functionality offered in the online Oracle Business Intelligence application.

F. ORACLE PUBLISHER

This component is an Enterprise Reporting and Distribution tool where reports designed for MS Word or Adobe Acrobat can be delivered via printer, e-mail, fax, webDAV or published to a portal. Figure 32 presents a set of reports provided by BI publisher from the very simple report to a more sophisticate printed and signed check.

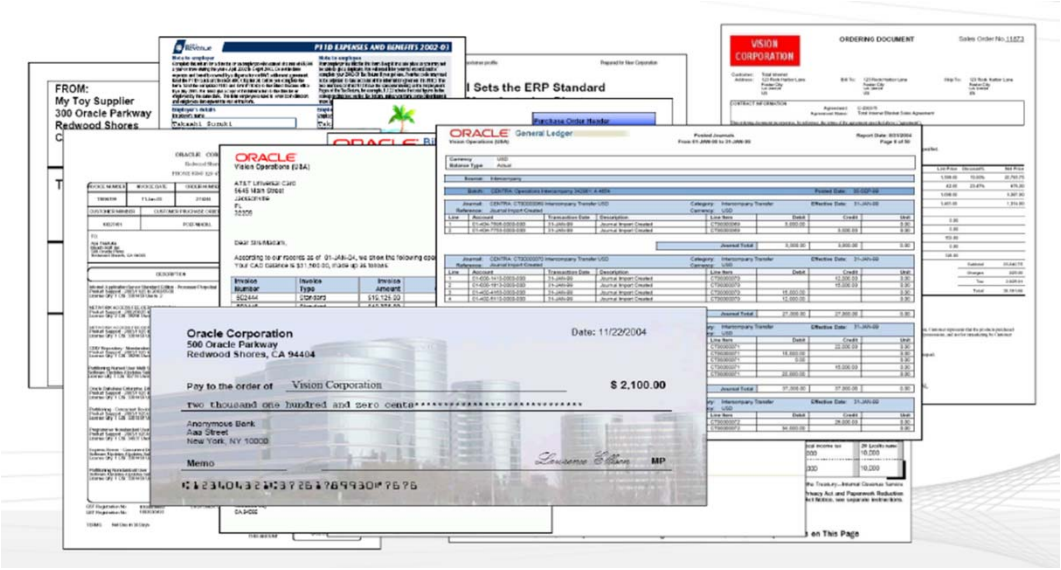


Figure 32 Oracle Publisher. (After [19])

BI publisher is a 100% thin-client application using WYSIWYG design environment. It provides “Pixel perfect” documents. BI publisher is interactive and easy to use similar to MS Office with instant preview. It performs OLAP and exploits

unstructured sources. Additionally, Oracle Publisher provides numerous formats for outputs such as XML, HTML, Word, PPT, PDF, RTF, etc.

G. REPORTING TOOLS

Interactive reporting is a Hyperion stack component. This tool provides executives, business users and analysts with user directed query and analysis means as well as interactive ad-hoc reporting.

SQR stands for Structured Query Reporter. SQR Production reporting is a Hyperion stack component. This module generates high volume, presentation-quality and pixel-perfect formatted reports with high performance regardless to the type and number of distinct sources.

Financial reporting, a Hyperion stack component, provides formatted financial and management reports that act in accordance with regulations and features currency translations, GAAP[2], IFRS[3] and other financial standards.

H. WEB ANALYSIS

This is a Hyperion stack components. It provides a web-based online analytical processing, presentation and reporting.

I. CONCLUSION

In conclusion, Oracle BI Tools provides solution for query, OLAP, reporting as well as scorecards. OBIEE conveys a full potential set of analytic and reporting. Yet, in order to perform data integration and transformation, other Oracle tools need to be installed along with OBIEE.

[2] GAAP: generally accepted accounting principles: a collection of rules and procedures and conventions that define accepted accounting practice

[3] IFRS: International Financial Reporting Standards (IFRS) are principles-based Standards, Interpretations and the Framework (1989) adopted by the International Accounting Standards Board (IASB).

THIS PAGE INTENTIONALLY LEFT BLANK

V RAPIDMINER DATA AND TEXT MINING SOFTWARE

This chapter describes the processing and analytics capabilities of RapidMiner, an open source Business Intelligence software product by Rapid-I company, for data import, data transformation, and data and text mining. Rapid-I supplies software, solutions, and services for data mining, text mining, as well as predictive analytics. Rapid-I products enable informed decisions and process optimization.

RapidMiner is an open-source solution for data and text mining. Depending on the license type and included extensions, Rapid-I offers a free community edition and three categories of enterprise edition: small, standard or developer edition [21]. It is available as a stand-alone version for data analysis or as part of the enterprise server system called RapidAnalytics. RapidAnalytics allows storing data and executing remote processes as well as advanced scheduling. In addition, it exposes RapidAnalytics Processes as Web Services and creates simple or interactive reports and dashboard elements.

In addition to RapidMiner data transformation and analysis solution, several other Rapid-I products are available. For instance, complex relationships and structures can easily be displayed, analyzed, and visually explored with RapidNet. Real-time market insights for customer and competitive intelligence are performed with RapidSentilyzer. BuzzBoard allows sentiment and opinion analysis [21]. Web service based automated document classification engine is assured by RapidDoc.

A. DESIGN PERSPECTIVE

This section presents the main components of the design perspective and how to create a process flow in Rapid Miner.

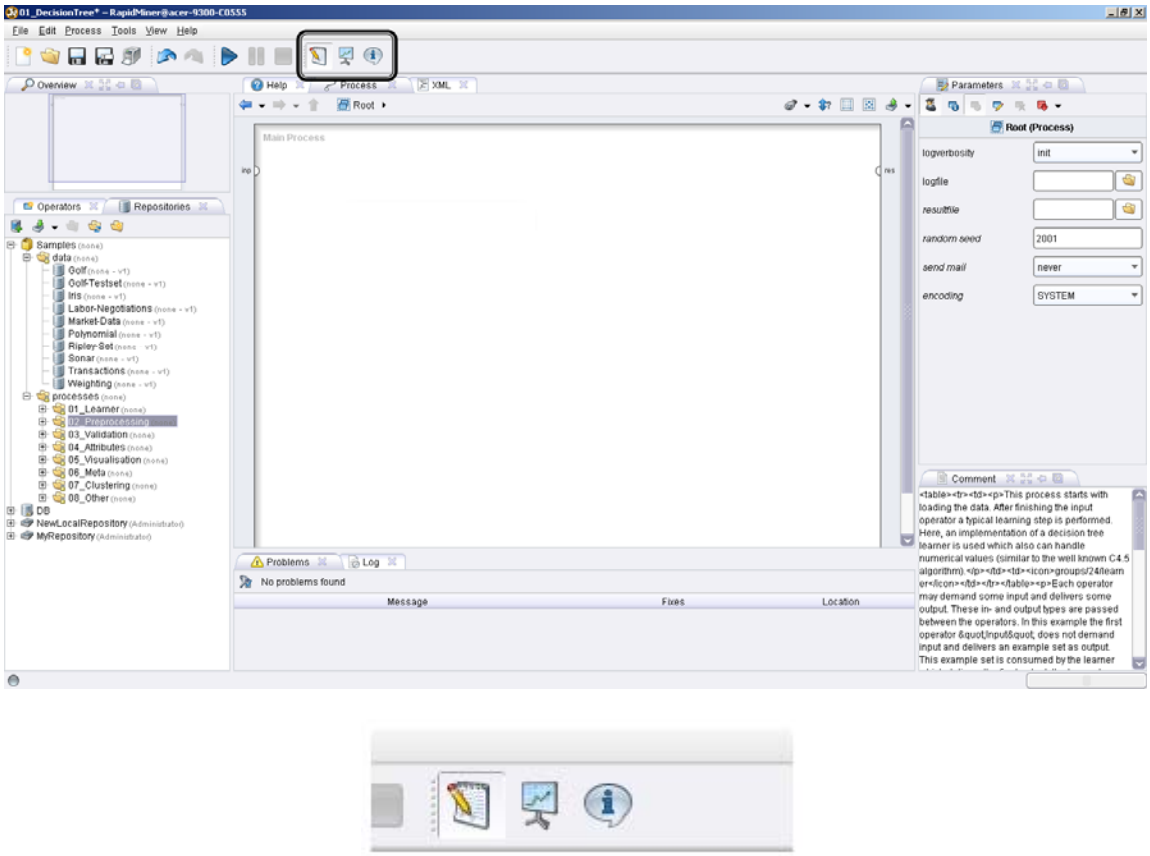


Figure 33 Toolbar icons for perspectives.

As shown in Figure 33, RapidMiner offers three perspectives accessible via the perspectives toolbar icons. The main perspective is the Design Perspective where all analysis processes are created and managed. The Result Perspective displays all data and models produced as results of a process. The Welcome Perspective is the initial view after starting the program.

1. Operators

Operators in RapidMiner are the main process components defining the analysis chain as a succession of entities in work steps. An operator is defined by several parameters such as the description of the expected inputs and the supplied outputs, the action applied on the input, and other parameters that control the performed action. As shown in Figure 34, the inputs and outputs of operators are generated or consumed via ports.

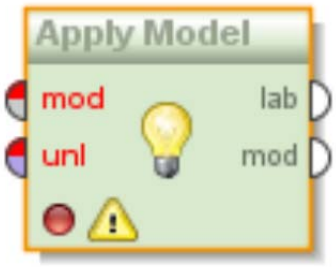


Figure 34 Operator connections, input ports versus output ports.

Rapid Miner contains more than 500 operators to support numerous tasks of data analysis. This includes input, output, extraction, transformation, loading, modeling and other aspects of data analysis.

An input operator can be used to import data from a repository, a database, or from a file. This type of operator does not require any input port. Instead, it has a parameter that specifies the source data location. Some operators, such as the data transform operator, transform their inputs to an object of similar type. Others, such as data mining methods transform the input to a different object type to deliver a model for the input data.

2. Processes

As illustrated in Figure 35, the process view illustrates a process consisting of several operators and their interconnections. A process can, for example, load data from a data source, transform the data, apply a data mining model, and export the results to a file. A process can consist of several hundred operators and be divided over several levels of subprocesses. The process illustrated in Figure 35 uses the frequency discretization operator to discretizes numerical attributes by putting the values into equal sized containers, and the nominal to binominal filter operator These operators are used in certain learning schemas that support special value types. The frequent item set mining operator FPGrowth for example supports only binominal features [22]. It is used to calculate item set often happening together.

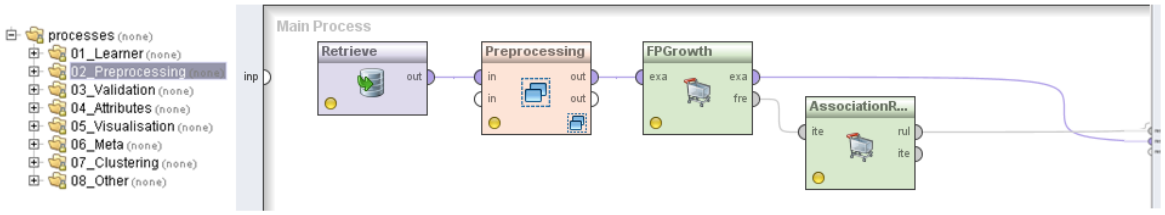


Figure 35 Processes in RapidMiner.

B. DATA IMPORT AND REPOSITORIES

This section describes data import features and capabilities of RapidMiner to import data from multiple data sources into the RapidMiner repository. As illustrated in Figure 36, a repository is a central storage location for all data, Metadata, and processes. This section will also discuss Metadata and its importance.

1. Importing Data and Objects into the Repository

RapidMiner offers several techniques to import data or models into the repository. These techniques include using Wizards, integrating the "Store" operator into the process, importing other formats by means of operators, or just storing objects from the Results and Process views.

Figure 36 Structure of data in the repository.

The wizard tool allows integrating data with different format and databases to the repository. This tool allows the user to simply drag and drop the file to be imported in the analysis process.

When using ETL in the process flow, it is possible to export the output directly to the repository using the “Store” operator. As shown in Figure 37, this operator has only one parameter, which specifies the repository location. Usually, it is utilized to perform automatic or a regular integration or transformation process.

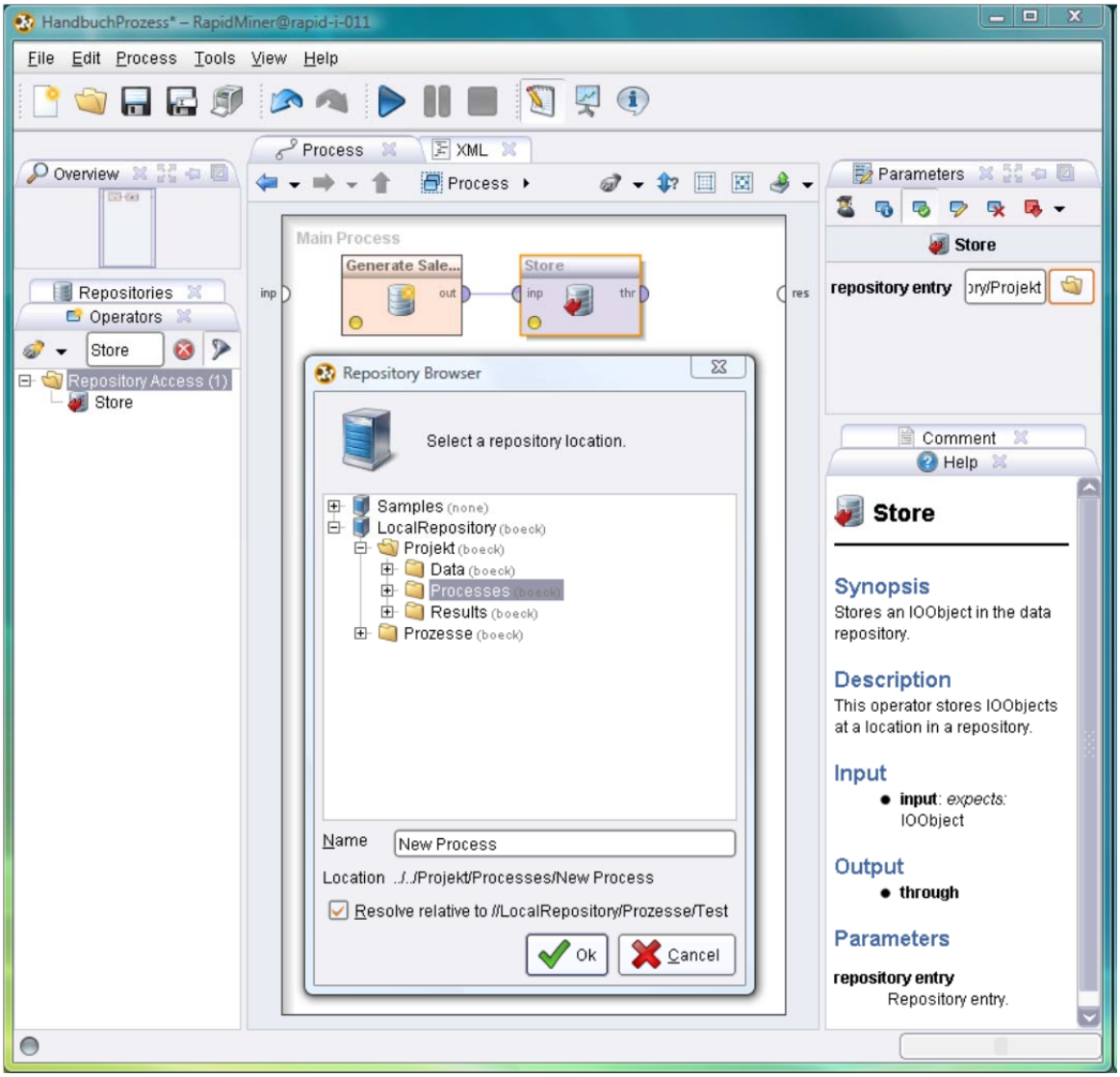


Figure 37 Using the Store operator to import data into the repository.

RapidMiner allows storing data and metadata in the repository as a dataset from different type of sources including CSV, Excel, SQL databases, etc. Although the use of import operators presents metadata availability problems, they are compulsory in the ETL process.

After the execution of a process, the result perspective allows storing the selected result directly in the repository.

2. Metadata

Whether the object is a model or dataset, metadata is generally defined as a description of the characteristics of a concept. Metadata in the context of dataset is defined as the information that describes data. Similar to data, metadata is also stored in the repository. RapidMiner allows managing data in the repository and consulting metadata as well. Figure 38, depicts the meta data of the output port of the operator "Discretize."

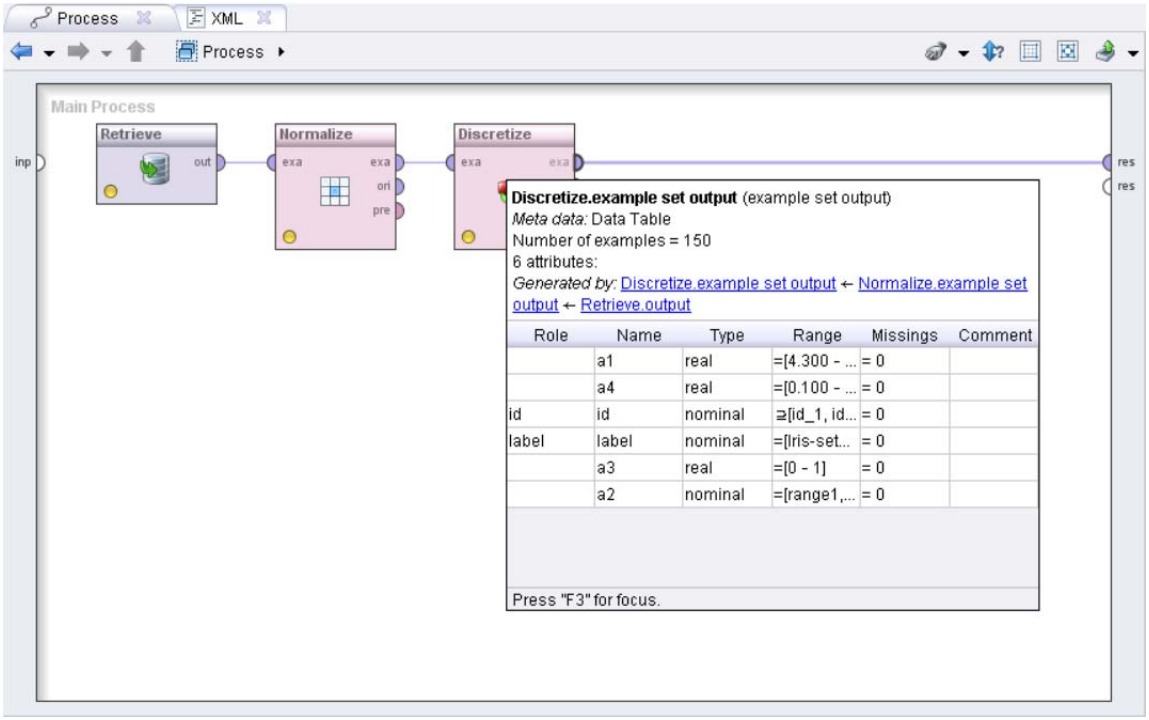


Figure 38 The Meta data of the output port of the operator "Discretize."

C. DATA TRANSFORMATION AND MODELING

Data Transformation is one of the most important functions in RapidMiner. Transformation operators are meant to process both data and metadata.

1. Basic Preprocessing Operators

RapidMiner offers numerous operators for basic preprocessing of data. This includes operators for sorting, filtering, aggregation, data cleansing, type conversion, etc.

These operators can perform operations to eliminate unneeded variables, transform variables, and create new variables in preparation for the data modeling step.

2. Modeling and Scoring

RapidMiner offers users many data modeling methods for a variety of tasks. This includes data mining, text mining, web mining, sentiment analysis, forums opinion mining, prediction, and model learning. A variety of algorithms is used by these operators. This includes regression, decision trees, neural nets, prediction, classification, clustering algorithms, etc. For example, a decision tree operator can be used to build a prediction model that can predict the value of target variables based on the values of predictor variables in a training historic data set. Models can be applied for scoring as well as predicting the values of target variables for different cases.

D. VISUALIZATION

This section discusses the visualization capabilities in RapidMiner. After building and running a process, RapidMiner displays the results in the Result Perspective in the form of file cards, as shown in Figure 7. The result perspective allows displaying several results in different views.

1. System Monitor

The system monitor informs the user about current memory in use. It shows the maximum memory available and the maximum memory usable.

2. Displaying Results

As shown in Figure 35, operators connected to the result port are object of Result Perspective. Figure 40 is an example of decision tree displayed in the Result perspective.

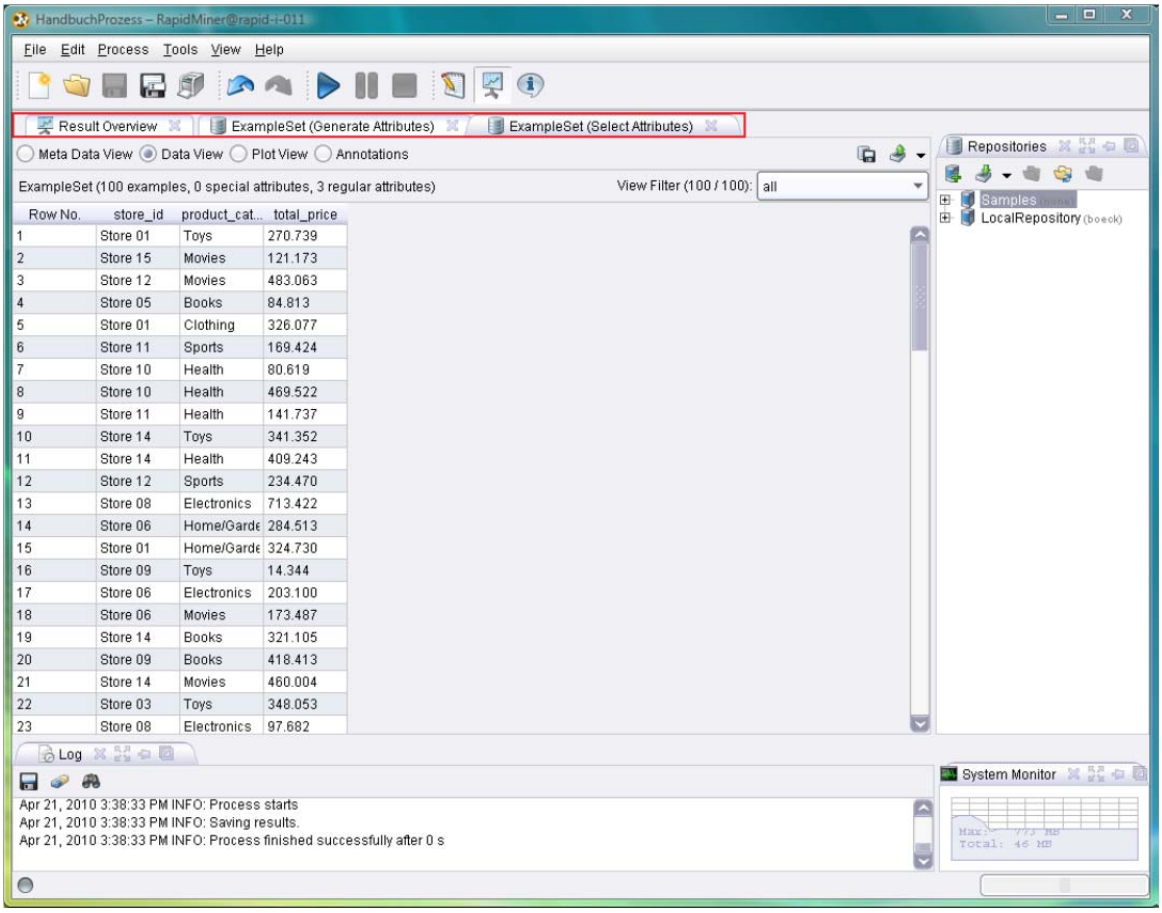


Figure 39 Result display.

As indicated in Figure 39, every displayed result or view is presented as an additional tab. In order to compare results, RapidMiner makes it possible to keep the old results. However, the user can manually close old results to avoid confusion.

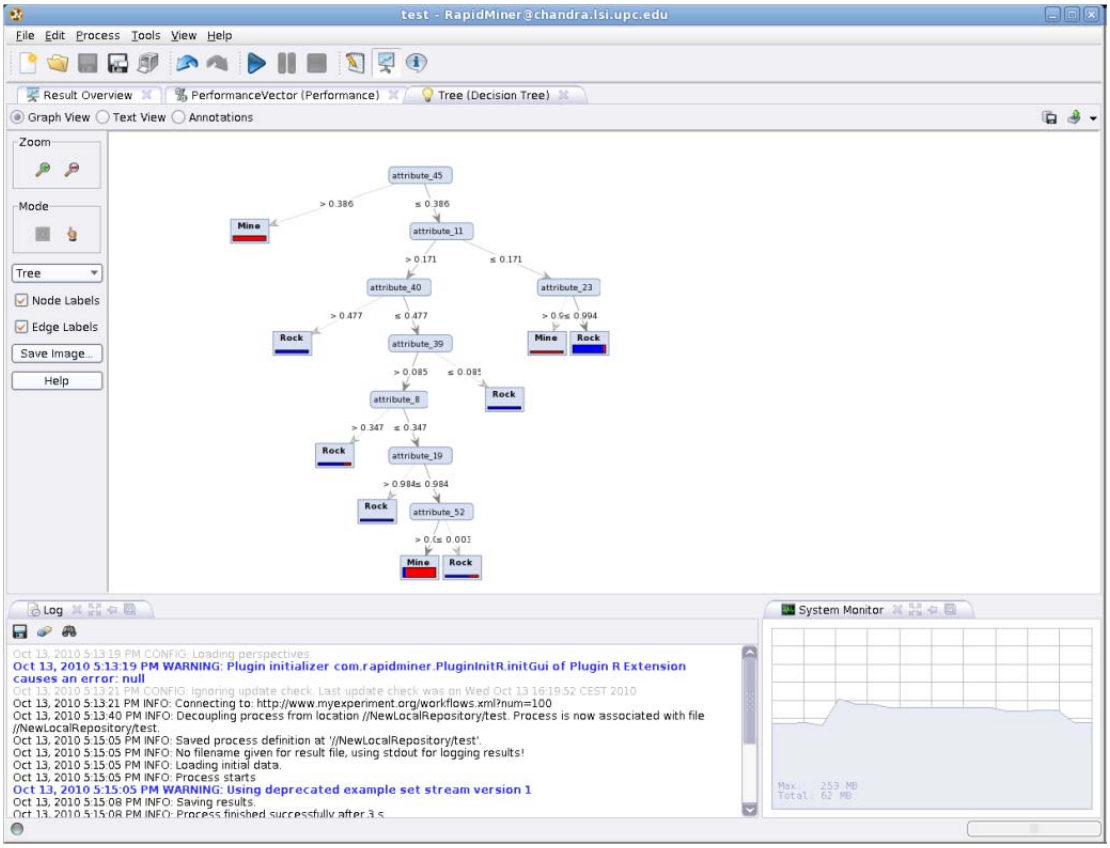


Figure 40 Result Perspective of RapidMiner: Decision Tree.

3. Sources for Displaying Results

In order to display results, the analyst can connect the object or any other breakpoint directly to the result port. This way, user can display all process results in the result perspective. Additionally, RapidMiner allows loading results directly from repositories which helps reviewing and comparing results. A third option, as shown in Figure 41, is to perform a transitional result right at the port output of an operator.

Figure 41 Display of results which are still at ports

4. Display Format

a. Text

The primary type of display is text format. Text view is used to visualize both models and results. Figure 42 shows an example of model results in textual form.

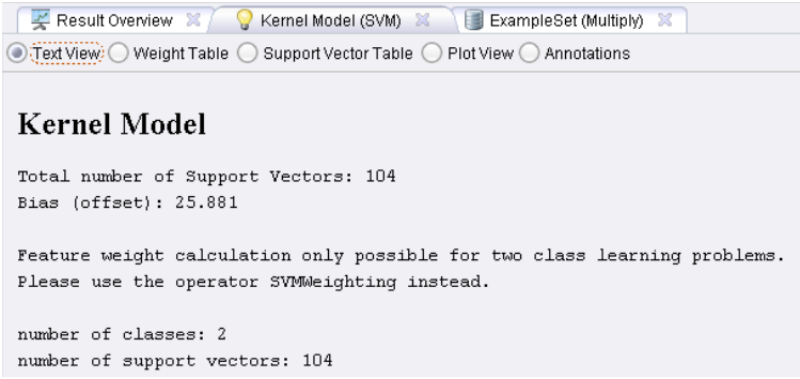
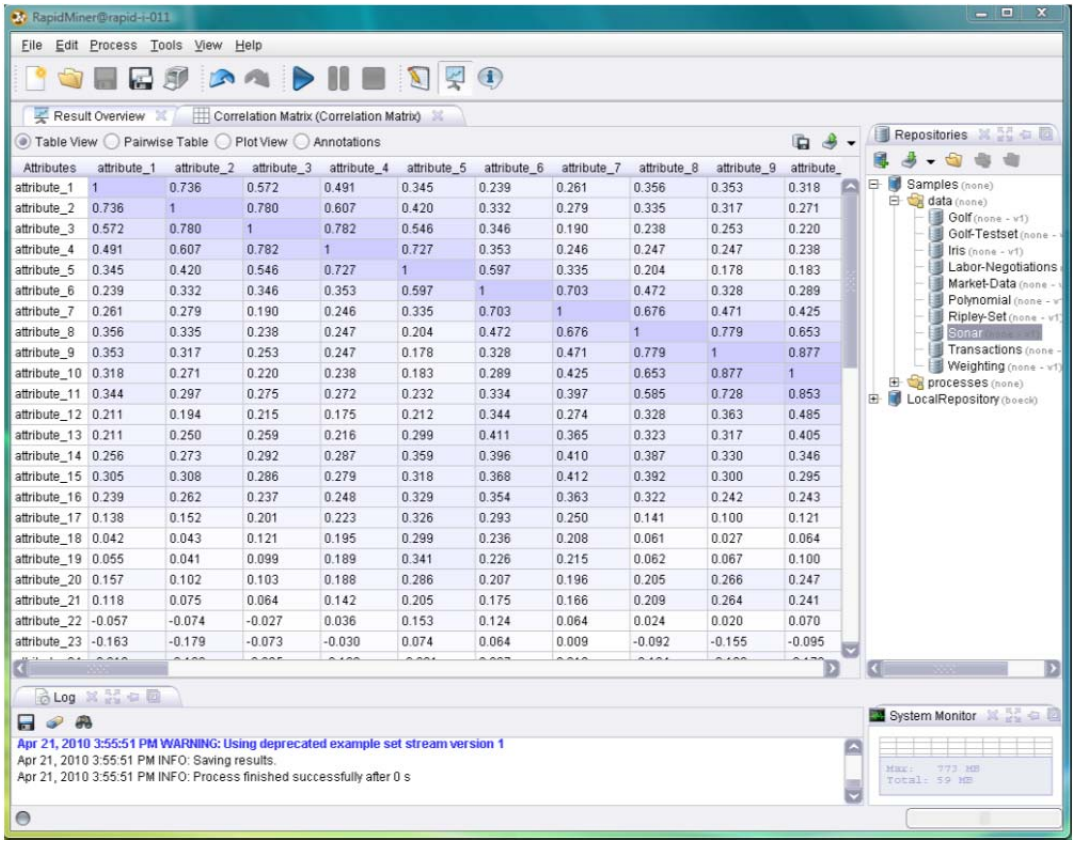


Figure 42 Example of Kernel Model displayed in text view

b. Tables

Results in RapidMiner can be also displayed in tabular format to show a data or metadata view. Moreover, as illustrated in Figure 43, table view can display a matrix showing how well inputs are correlated.



Attributes	attribute_1	attribute_2	attribute_3	attribute_4	attribute_5	attribute_6	attribute_7	attribute_8	attribute_9	attribute_
attribute_1	1	0.736	0.572	0.491	0.345	0.239	0.261	0.356	0.353	0.318
attribute_2	0.736	1	0.780	0.607	0.420	0.332	0.279	0.335	0.317	0.271
attribute_3	0.572	0.780	1	0.782	0.546	0.346	0.190	0.238	0.253	0.220
attribute_4	0.491	0.607	0.782	1	0.727	0.353	0.246	0.247	0.247	0.238
attribute_5	0.345	0.420	0.546	0.727	1	0.597	0.335	0.204	0.178	0.183
attribute_6	0.239	0.332	0.346	0.353	0.597	1	0.703	0.472	0.328	0.289
attribute_7	0.261	0.279	0.190	0.246	0.335	0.703	1	0.676	0.471	0.425
attribute_8	0.356	0.335	0.238	0.247	0.204	0.472	0.676	1	0.779	0.653
attribute_9	0.353	0.317	0.253	0.247	0.178	0.328	0.471	0.779	1	0.877
attribute_10	0.318	0.271	0.220	0.238	0.183	0.289	0.425	0.653	0.877	1
attribute_11	0.344	0.297	0.275	0.272	0.232	0.334	0.397	0.585	0.728	0.853
attribute_12	0.211	0.194	0.215	0.175	0.212	0.344	0.274	0.328	0.363	0.485
attribute_13	0.211	0.250	0.259	0.216	0.299	0.411	0.365	0.323	0.317	0.405
attribute_14	0.256	0.273	0.292	0.287	0.359	0.396	0.410	0.387	0.330	0.346
attribute_15	0.305	0.308	0.286	0.279	0.318	0.368	0.412	0.392	0.300	0.295
attribute_16	0.239	0.262	0.237	0.248	0.329	0.354	0.363	0.322	0.242	0.243
attribute_17	0.138	0.152	0.201	0.223	0.326	0.293	0.250	0.141	0.100	0.121
attribute_18	0.042	0.043	0.121	0.195	0.299	0.236	0.208	0.061	0.027	0.064
attribute_19	0.055	0.041	0.099	0.189	0.341	0.226	0.215	0.062	0.067	0.100
attribute_20	0.157	0.102	0.103	0.188	0.286	0.207	0.196	0.205	0.266	0.247
attribute_21	0.118	0.075	0.064	0.142	0.205	0.175	0.166	0.209	0.264	0.241
attribute_22	-0.057	-0.074	-0.027	0.036	0.153	0.124	0.064	0.024	0.020	0.070
attribute_23	-0.163	-0.179	-0.073	-0.030	0.074	0.064	0.009	-0.092	-0.155	-0.095

Figure 43 Correlated data displayed in table view.

c. Plots

A plot view allows a variety of visualization to illustrate data, models, or results.

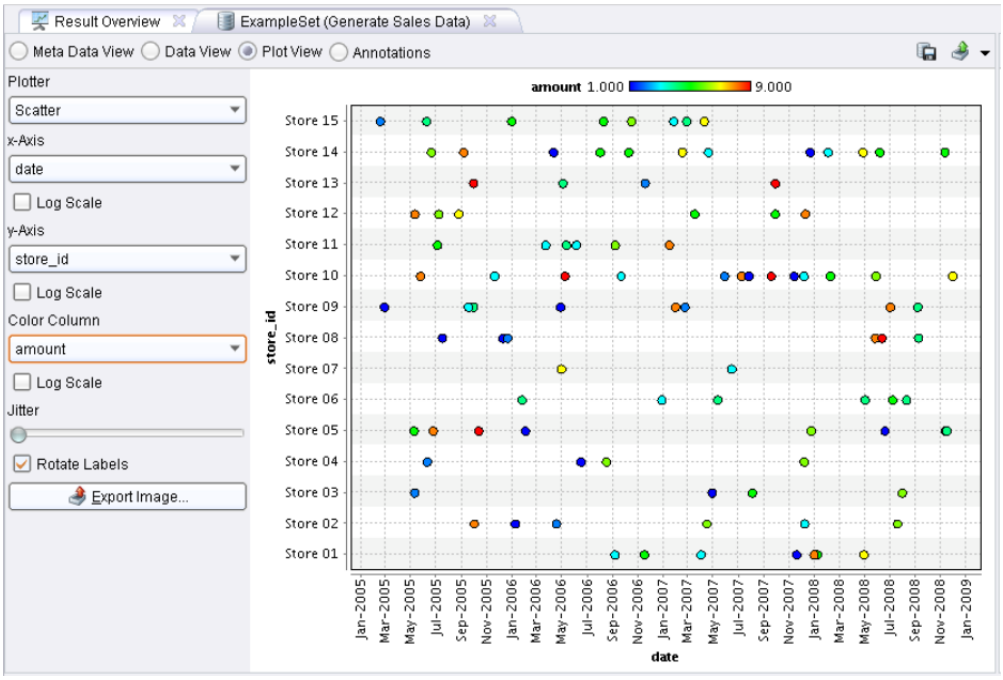


Figure 44 Visualization of a data set in a plot view.

Plot view offers around thirty methods for displaying data in 2D, 3D, and N dimensional plots. Figure 44 shows an example of scatter plot. Changing the plotter configuration defines how the view will be displayed. Figure 45 illustrates an example of bars stacked plot.

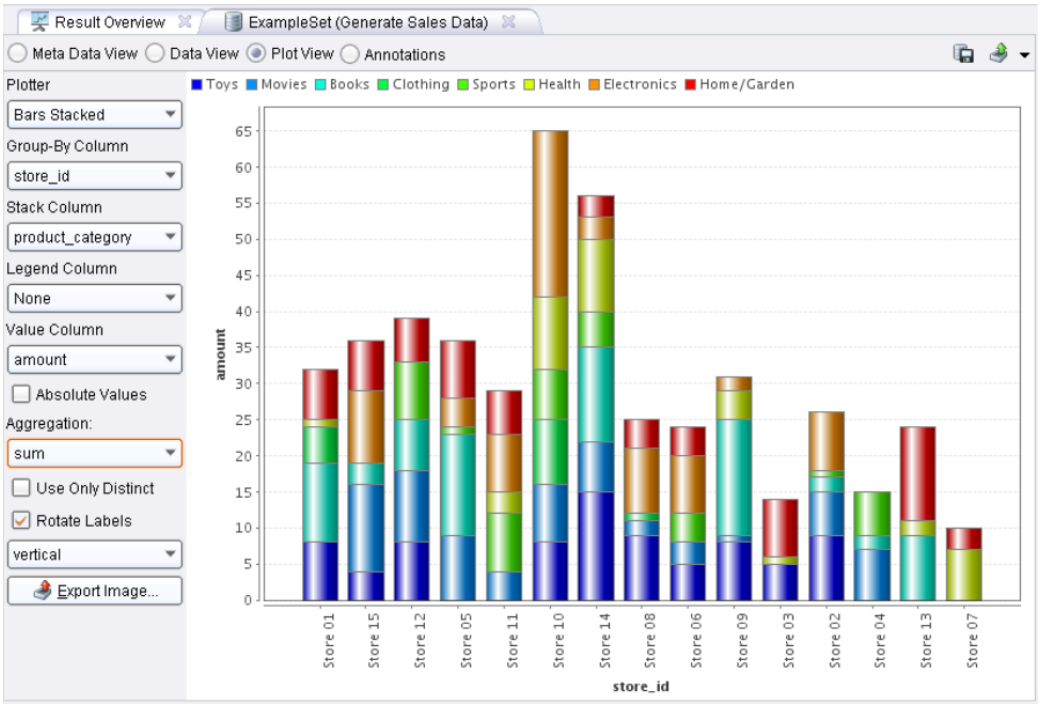


Figure 45 Example of bars stacked plot.

d. Graphs

In order to show the relationships between nodes, RapidMiner offers Graphs visualization. Figure 8 illustrates a decision tree represented as a hierarchical graph.

5. Validation and Performance Measurement

As shown in Figure 46, RapidMiner offers several evaluation tools. Evaluating the quality of a classifier is usually hard to perform because it depends on the size and the quality of the training dataset, irrelevant attributes, missing values, and so on. The easiest method to process a large dataset is by dividing it to a training set and a test set. However, it has been shown that this technique reports a more optimistic result than in reality. Numerous classifiers provide significant different results when trained with slightly changed training set.

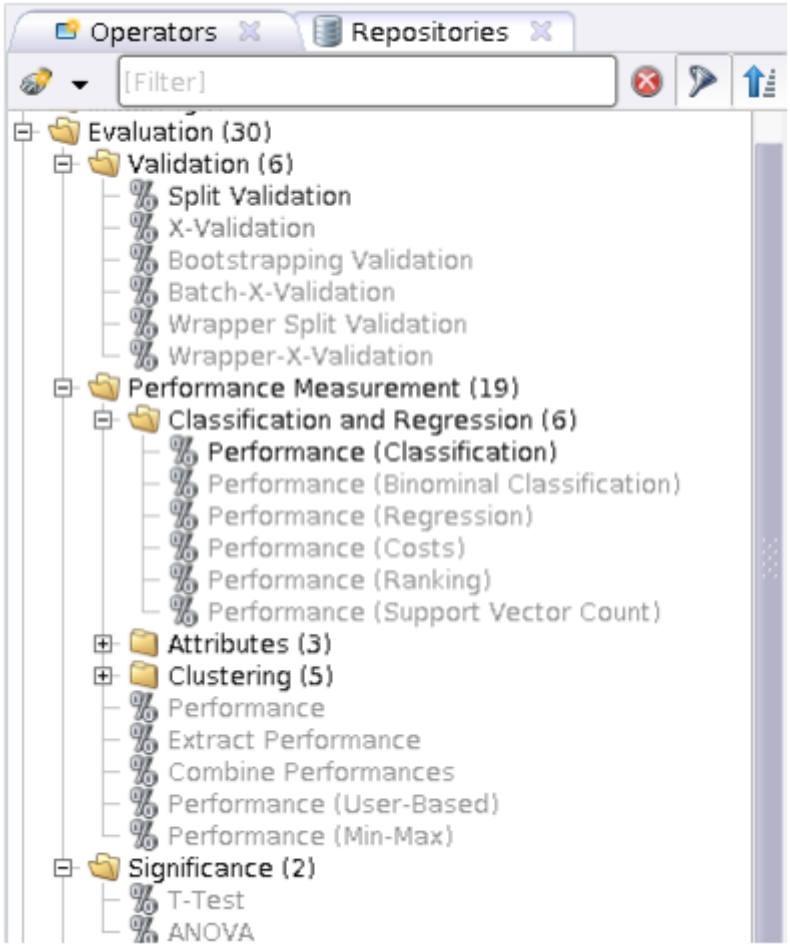


Figure 46 Evaluation tools

E. CONCLUSION

The open source analysis solution RapidMiner is an excellent tool for organizations looking for a free or low cost solution to conduct BI. It is available as a stand-alone version or within the enterprise server system called RapidAnalytics. It provides a range of data analysis and visualization in addition to data and text mining capabilities. RapidMiner offers a rich and complete set of tools helping with data integration, data transformation, data processing, machine learning, and evaluation.

The following table compares the capabilities of the community edition with that of the different versions of the enterprise edition. This table would help users deciding which version is needed to satisfy their BI needs.

	Community Edition	Enterprise Edition		
		Small	Standard	Developer
Number of Users	**1 Analyst**	**1 Analyst**	**5 Analysts**	**OEM Agreement**
Extension Packs for Increased Number of Users	✗	✓	✓	✓
License	Open Source	Open Source or Closed Source	Open Source or Closed Source	Open Source or Closed Source
Certified	✗	✓	✓	✓
Integration				
Into Open-Source Software	✓	✓	✓	✓
Into Closed-Source Software	✗	✗	✗	✓
Into Web Services	✗	✗	✗	✓
Guarantees				
Guarantee for Bugfixes	✗	✓	✓	✓
Intellectual Property Indemnification	✗	✓	✓	✓
Warranty for Services	✗	✓	✓	✓
Problem Resolution Support				
Community Forums	✓	✓	✓	✓
Community Web Documentation (Wiki)	✓	✓	✓	✓
Service Level Agreement	✗	✗	✓	✓
Number of Incidents	✗	✗	**Unlimited**	**Unlimited**
Web-based Case Management	✗	✗	✓	✓
Mail Support	✗	✗	✓	✓
Support Access	✗	✗	Business Hours	Business Hours

Maximum Initial Response Time	✗	✗	4 hours	4 hours
Emergency Hot Fix Build	✗	✗	✓	✓
Consultative Support				
Included Amount	✗	✗	8 Hours	8 Hours
Discount on Consulting Rates	✗	✗	25%	25%
Discount on In-House Training Rates	✗	✗	25%	25%
Remote Troubleshooting	✗	✗	✓	✓
Process Review	✗	✗	✓	✓
Process Optimization	✗	✗	✓	✓
Performance Tuning	✗	✗	✓	✓
Customer Code Review	✗	✗	✗	✓
Maintenance				
Software Maintenance	**By In-house Staff**	**By Rapid-I Engineers**	**By Rapid-I Engineers**	**By Rapid-I Engineers**
Updates via Update and Installation Server	✓	✓	✓	✓
Software Installation for Extensions	✓	✓	✓	✓
Patch Releases	✗	✓	✓	✓
Fixes Included in Future Releases	✗	✓	✓	✓
Stabilized and Certified Software Releases	✗	✓	✓	✓
Managed Release Cycles	✗	✓	✓	✓

Table 1 RapidMiner community vs. enterprise edition comparison (From [21])

THIS PAGE INTENTIONALLY LEFT BLANK

VI. SUMMARY, CONCLUSION, AND RECOMMENDATIONS

This chapter summarizes the effort of the thesis, provides conclusions, and suggests possible recommendations for a better BI tool selection.

A. SUMMARY

This thesis deals with the application and survey of business intelligence tools within the context of military decision making. It began, in Chapter II, by introducing Business Intelligence concepts and components. It described the data warehousing concept, architecture, and process. It then presented a detailed review of business analytics including the Online Analytic Process (OLAP). It then introduced data, text and web Mining techniques and tools and their applications to support organizational processes. The chapter concluded with an overview of performance measurement systems. Chapter III introduced Megaputer PolyAnalyst, a mainstream data and text mining tool. Following a brief overview of data mining concepts, the chapter discussed the integration, and data manipulation processes used by PolyAnalyst. It then presented a variety of machine learning algorithms used by PolyAnalyst including clustering, classification, decision tree, linear regression, link analysis, and text mining. The chapter concluded with an overview of PolyAnalyst reporting capabilities. The fourth chapter described a high level querying and reporting tool called Oracle Business Intelligence Enterprise Edition. It started by discussing the architecture and main components of the tool. This was followed by a detailed description of each component, which included BI answers, BI interactive dashboard, and its main components, segmentation and list generation, and disconnected analytics to support people disconnected from the Corporate Network. A discussion of Oracle Publisher and the enterprise reporting and distribution tool followed. The chapter concluded with a brief discussion of interactive reporting, SQR production reporting, financial reporting, and web analysis. The fifth chapter overviewed the processing and analysis capabilities of RapidMiner, an open source Business Intelligence software developed by Rapid-I company. It discussed in some detail its data transformation, data mining, and machine learning features as well as

its design perspective, the repository, data transformation, and visualization capabilities. It is hoped that the analysis of the different BI tools presented in this will help decision making teams in selecting the most appropriate BI tool to fulfill an organization needs.

B. CONCLUSION AND RECOMMENDATIONS

We make the following conclusions on the capabilities of tools studied in this research. Megaputer PolyAnalyst is a powerful data and text mining tool that provides the analyst with several capabilities to discover unseen relationships in data and text. It includes a comprehensive set of tools for data access from a variety of sources, data integration, data transformation, as well as exporting the results. Oracle BI Tools on the other hand, provides solution for query, OLAP, reporting as well as scorecards. Yet, in order to perform data integration and transformation, other tools need to be installed along with OBIEE. Similar to Megaputer PolyAnalyst, the open source analysis solution RapidMiner provides a range of tools data and text mining capabilities. It offers a complete set of tools for data integration, data processing, and machine learning.

In addition to the tool specific conclusions, the research of this thesis leads to the following observations/conclusions:

1. BI Means Different Things for Different People

BI is an umbrella term that involves the analysis of data using statistical and mathematical techniques. This includes a wide range of analysis such as querying and reporting, multidimensional analysis, data and text mining, mathematical programming, and calculating performance metrics for inclusion on dashboards/scorecards. These are different types of analysis with important implications. First, the type of problem needs to be matched properly with the analysis technique chosen. Second, the right people with the right set of skills need also to match the type of analysis used. Finally, the technologies used need to be chosen carefully to support the type of analysis chosen.

2. BI is becoming a Critical Requirement for Organizations

For many organizations, including military ones, BI is evolving from a “nice-to-have” application to a critical organizational requirement for collecting, storing,

analyzing, and providing access to data to help users make better and faster decisions. BI also becomes a critical approach to analyze current business processes and to design better ones. Additionally, for many business processes, it is becoming imperative to integrate BI solutions into work flows to monitor and increase their efficiency and effectiveness.

3. Big Data is changing the Scope and Technologies for BI

Most medium and large organization collect, store, and analyze "big data." In addition to the structured data of operational systems, organizations are capturing and storing text data from their websites, call centers, surveys, e-mail, documents, social media, etc. There are more data sources, and the data is arriving at a higher velocity. This vast amount of structured and unstructured data contains a wealth of potentially useful information but creates challenges for capturing, storing, and analyzing it. It is therefore imperative for organizations to plan for and integrate big data into their BI strategy, architecture, technologies, processes, and activities. Failure to do so would result in a new generation of analytic silos similar to the data silos of the 70s and 80s.

4. BI is used in Novel Applications

Most people do not think of many organizational processes as being potential for BI. It is important to think carefully and deeply how BI can improve decision making for nontraditional application areas. It is likely that the highest return on investment would be obtained from such applications.

5. BI Requires a Wide and Varied Set of Skills

Analysts who perform analytics must have a wide and varied set of skills: the ability to work with large data sets, an understanding of analysis methods, domain knowledge, and communications skills. Few people are strong in all of these areas. An analyst may not possess all of these skills, but someone on the team must. Organizations must also be prepared to develop internal, analytics oriented training programs to grow necessary skills.

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF REFERENCES

[1] Simons, R. (2002). *Performance Measurement and ControlSystems for Implementing Strategy.* Upper Saddle River, NJ: Prentice Hall

[2] Efraim Turban, Ramesh Sharda, Jay E. Aronson, David King (2008). *Business Intelligence: A managerial Approach.* Upper Saddle River, New Jersey: Pearson Education, Inc.

[3] Informationweek.com . "U.S. military using business intelligence to track wounded soldiers."Available: http://www.informationweek.com/news/healthcare/admin-systems/216500651

[4] MicroStrategy.com. "MicroStrategy business intelligence and mobile business intelligence." Available: http://www.microstrategy.com

[5] Ekerson,W. (2003). *Smart Companies in the21st Century: the secret of creating successful business intelligence solutions*. Seattle, WA: The datawarehousing institute.

[6] Dundas.com. "Dundas software" Available: dundas.com/dashboard/online-examples/screenshots/Technology-Performance-Dashboard.aspx [Accessed 20 December 2011].

[7] ZSL Inc. "Dashboards vs scorecards." Available: http://www.performancesolutions.nc.gov/metrics/DevelopingInsightsFromMetrics/docs/DashboardScorecards.pdf [Accessed Mai 2012].

[8] Microstrategy.com. "Micro Strategy" Available: http://www.microstrategy.com/QuickTours/HTML/CompIntel_Template/QuickTour/index.asp?vendor=SAS&page=2 [Accessed 20 December 2011].

[9] Wayne W. Eckerson (2006).*Performance Dashboards: Measuring, Monitoring, And Managing Your Business* . Hoboken, NJ: John Wiley & Sons, Inc.

[10] Joseph Raynus (2011). *Quantitative Business Performance Management: Challenge, Change, and Dashboard.* Taylor & Francis Group, Boca Raton FL 33487–2742

[11] Anderson.ucla.edu "Data mining: What is data mining?" Available: http://www.anderson.ucla.edu/faculty/jason.frand/teacher/technologies/palace/datamining.htm [Accessed May 2012]

[12] "Chapter 2 Data warehousing." Available: http://www70.homepage.villanova.edu/matthew.liberatore/Mgt2206/CH02.ppt_2206.ppt [Accessed May 2012]

[13] Christian S. Jensen, Torben Bach Pedersen, Christian Thomsen (2010). "*Multidimensional Databases and Data Warehousing*." San Francisco, CA: Morgan & Claypool Publishers.

[14] Ralph Kimball, Margy Ross (2010). "*The Kimball Group Reader: Relentlessly Practical Tools for Data Warehousing and Business Intelligence*." Indianapolis, IN: Wiley Publishing, Inc..

[15] Megaputer.com "Megaputer intelligence Inc. (2007): PolyAnalyst 6" Available: http://www.megaputer.com/down/PolyAnalyst_6_brochure.pdf [Accessed October 2011]

[16] Hand, D. J., H. Mannila, and P. Smyth (2001). *Principles of Data Mining. Adaptive Computation and Machine Learning*. Cambridge MA: MIT Press.

[17] Gregory Piatetsky-Shapiro, (February 1997). "Data mining community's top resource," Available: http://www.kdnuggets.com/faq/index.html [Accessed October 2011]

[18] PolyAnalyst online help complied HTML (*.chm*) format. Available only with the installation: "C:\Program Files\Megaputer Intelligence\PolyAnalyst 6.0\Help\help.chm"

[19] ncoaug.org. "OBIEE 11g new features" Available: http://www.ncoaug.org/NCOAUG%20Training%20Day%20Feb%202011/OB_11g_New_Features_v3.pdf [Accessed 10 October 2011]

[20] Oracle.com. "Oracle VM template for oracle business intelligence enterprise edition (OBIEE)" Available: http://www.oracle.com/technetwork/server-storage/vm/obiee-092989.html [Accessed 5 September 2011].

[21] Rapid-i.com. "Rapid miner official website" Available: http://www. rapid-i.com [Accessed November 2011]

[22] Rapid-I GmbH (2010): "Rapidminer-5.0 manual v1.0." Available: http://rapid-i.com/content/view/26/84/ [Accessed 10 February 2012]

INITIAL DISTRIBUTION LIST

1. Defense Technical Information Center
   Ft. Belvoir, Virginia

2. Dudley Knox Library
   Naval Postgraduate School
   Monterey, California

3. Dr. Magdi Kamel Ph.D.
   Information Sciences Department
   Naval Postgraduate School
   Monterey, California

4. Dr. Walter Kendall
   Information Sciences Department
   Naval Postgraduate School
   Monterey, California

5. Mr. Arijit Das
   Information Sciences Department
   Naval Postgraduate School
   Monterey, California

6. Capt. Mohamed Ilyes Tounsi
   Tunis, Tunisia

7. D.C. Boger
   Information Sciences Department
   Naval Postgraduate School
   Monterey, California